DOCUMENT RESUME

ED 384 474 RC 020 177

TITLE	Science and Math Assessment in K-6 Rural and Small Schools. Rural, Small Schools Network Information Exchange: Number 14, Spring 1993.
INSTITUTION	Regional Laboratory for Educational Improvement of the Northeast & Islands, Andover, MA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	93
CONTRACT	RP91002008
NOTE	185p.; Photographs will not reproduce adequately.
PUB TYPE	Collected Works - General (020)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	*Academic Standards; Curriculum Based Assessment; Educational Change; Educational Practices; Educational Testing; Elementary Education; *Elementary School Mathematics; *Elementary School Science; Evaluation Methods; Informal Assessment; *Portfolio Assessment; Program Descriptions; *Rural Education; *Student Evaluation
IDENTIFIERS	*Alternative Assessment; Authentic Assessment

ABSTRACT

This packet includes reprints of journal articles and other resources concerning the assessment of science and math in small, rural elementary schools. Articles include: (1) "Standards, Assessment, and Educational Quality" (Lauren B. Resnick); (2) "A True Test: Toward More Authentic and Equitable Assessment" (Grant Wiggins); (3) "How World-Class Standards Will Change Us" (Arthur L. Costa); (4) "Smart Tests" (Deborah L. Cohen); (5) "Laser Disk Portfolios: Total Child Assessment" (Jo Campbell); (6) "Portfolios Invite Reflection--from Students and Staff" (Elizabeth A. Hebert); (7) "Portfolio Assessment in the Hands of Teachers" (Clare Forseth); (8) "Portfolio Assessment" (Susan Black); (9) "Assessing the Outcomes of Computer-Based Instruction: The Experience of Maryland" (Gita Z. Wilder, Mary Fowles); (10) "Why Standards May Not Improve Schools" (Elliot W. Eisner); (11) "Assessing Alternative Assessment" (Gene I. Maeroff); (12) "Assessment Recordkeeping in a Non-Graded Developmentally-Based Program" (Elsbeth Bellemere, Jeanne King); (13) "Strategies for the Development of Effective Performance Exercises" (Joan Boykoff Baron); (14) "Evaluating Elementary Science" (Rodney L. Doran and others); (15) "Science for All: Getting It Right for the 21st Century" (Kenneth M. Hoffman, Elizabeth K. Stage); (16) "Active Assessment for Active Science" (George E. Hein); (17) "The Nature of Elementary Science: What Does 'It' Look Like?" (Gregg Humphrey); (18) "Assessment: What Is 'IT'?" (Gregg Humphrey); (19) "What's Worth Assessing?" (Monte Moses); (20) "Creating Benchmarks for Science Education" (Andrew Ahlgren); (21) "Assessment, Practically Speaking" (Lehman W. Barnes, Marianne B. Barnes); (22) "Getting Connected to Science" (Candace L. Julyan); (23) "EDTALK: What We Know about Science Teaching and Learning"; (24) "What We've Learned about Assessing Hands-On Science" (Richard J. Shavelson, Gail P. Baxter); (25) "NCTM's Standards: A Rallying Flag for Mathematics Teachers" (Thomas A. Romberg); (26) "Measuring What's Worth Learning"; (27) "Report Offers Glimpse of Mathematics Assessment of the Future" (Robert Rothman); (28) "The Power of Thinking Mathematics" (Alice J. Gill, Lovely H. Billups); (29) "Bringing Meaning to Math with a Student-Run Store" (Deborah Black); (30) "Employer Expectations for School Mathematics" (Henry O. Pollak); and (31) "Evaluating Problem Solving in Mathematics" (Walter Szetela, Cynthia Nicol). (LP)

ED 384 474

RURAL, SMALL SCHOOLS NETWORK
INFORMATION EXCHANGE

Number 14

SPRING 1993

SCIENCE AND MATH ASSESSMENT IN K-6 RURAL AND SMALL SCHOOLS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Janet Angelis
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Regional Laboratory
for Educational Improvement of the Northeast & Islands

BEST COPY AVAILABLE

This publication is based on work sponsored wholly or in part by the U.S. Department of Education under contract number RP 91002008. The content of the publication does not necessarily reflect the views of the department or any other agency of the U.S. Government.

RC020177

The Regional Laboratory
for Educational Improvement of the Northeast & Islands

Spring 1993

Dear Rural, Small School Leader:

The need for alternative assessment methods is evolving alongside a growing shift from schools that are content-centered to schools that are learner-centered. Higher order thinking processes, decision making skills and cooperative grouping strategies are just a few of the myriad student proficiencies that schools are choosing to assess. The range of articles in this Rural, Small Schools Network Information Packet includes a discussion of assessment standards, an overview of a variety of assessment methods, and descriptions of 'state of the art' science and math programs.

Students are assessed on what they learn; as well as how they respond to a test. In the past educators have relied heavily on one testing instrument. Multi-faceted assessment instruments like portfolios, essays, multiple choice questions, and hands-on presentation instruments now provide a more in-depth perspective of what students have learned as well as offering students a variety of ways to demonstrate their knowledge.

We hope you find this packet informative as you look at assessment as a way of gathering information about students' learning processes. An evaluation card has been provided so you can send us your feedback. We also welcome your suggestions for future Information Exchange Packet topics. Please jot any ideas you may have on the card or contact us at the Rural, Small Schools Network, 83 Boston Post Road, Sudbury, MA 01776, (508) 443-7991.

Sincerely,

John R. Sullivan, Jr., Ed.D.
Program Director
Rural, Small Schools Network

Virginia L. Warn
Associate Program Director
Rural, Small Schools Network

RCO 20177

CONTENTS

SCIENCE AND MATH ASSESSMENT IN K-6 RURAL AND SMALL SCHOOLS

SECTION I: Overview of Assessment

"Standards, Assessment, and Educational Quality," by Lauren B. Resnick, Stanford Law and Policy Review, Winter 1992-93.

"A True Test: Toward More Authentic and Equitable Assessment," by Grant Wiggins, Phi Delta Kappan, May 1989.

"How World-Class Standards Will Change Us," by Arthur L. Costa, Educational Leadership, February 1993.

"Smart Tests," by Deborah L. Cohen, Teacher Magazine, March 1993.

"Laser Disk Portfolios: Total Child Assessment," by Jo Campbell, Educational Leadership, May 1992.

"Portfolios Invite Reflection--from Students and Staff," by Elizabeth A. Hebert, Educational Leadership, May 1992.

"Portfolio Assessment in the Hands of Teachers," by Clare Forseth, The School Administrator, December 1992.

"Portfolio Assessment," by Susan Black, The Executive Educator, February 1993.

"Assessing the Outcomes of Computer-Based Instruction: The Experience of Maryland," by Dr. Gita Z. Wilder and Mary Fowles, T.H.E. Journal, September 1992.

"Why Standards May Not Improve Schools," by Elliot W. Eisner, Educational Leadership, February 1993.

"Assessing Alternative Assessment," by Gene I. Maeroff, Phi Delta Kappan, December 1991.

"Assessment Recordkeeping in a Non-Graded Developmentally-Based Program," by Elsbeth Bellemere and Jeanne King, presented at the High/Scope Foundation Conference on Assessment in Ypsilanti, Michigan, Winter 1990.

"Strategies for the Development of Effective Performance Exercises," by Joan Boykoff Baron, Applied Measurement in Education, copyright 1991.

SECTION II: Science Assessment K-6

"Evaluating Elementary Science," by Rodney L. Doran, Douglas Reynolds, Janice Camplin, and Nicholas Hejaily, Science and Children, November/ December 1992.

"Science for All: Getting It Right For the 21st Century," by Kenneth M. Hoffman and Elizabeth K. Stage, Educational Leadership, February 1993.

"Active Assessment for Active Science," by George E. Hein, Expanding Student Assessment, edited by Vito Perrone, copyright 1991 by the Association for Supervision and Curriculum Development.

"The nature of elementary science: what does "it" look like?," by Gregg Humphrey, Teachers' Lab, September 1991.

"Assessment: what is "IT?," by Gregg Humphrey, Teachers' Lab, November & Decemher 1991.

"What's Worth Assessing?," by Monte Moses, The School Administrator, December 1992.

"Creating Benchmarks For Science Education," by Andrew Ahlgren, Educational Leadership, February 1993.

"Assessment, Practically Speaking," by Lehman W. Barnes and Marianne B. Barnes, Science and Children, March 1991.

"Getting Connected to Science," by Candace L. Julyan, Hands On!, Volume 14, Number 1, Spring 1991.

EDTALK: What We Know About Science Teaching And Learning, published by the Council for Educational Development and Research, pages 57-62.

"What We've Learned About Assessing Hands-On Science," by Richard J. Shavelson and Gail P. Baxter, Educational Leadership, May 1992.

SECTION III: Math Assessment K-6

"NCTM's Standards: A Rallying Flag For Mathematics Teachers," by Thomas A. Romberg, Educational Leadership, February 1993.

"Measuring What's Worth Learning," a report from National Academy Press, copyright 1993.

"Report Offers Glimpse of Mathematics Assessment of the Future," by Robert Rothman, Education Week, December 9, 1992.

"The Power of Thinking Mathematics," by Alice J. Gill and Lovely H. Billups, American Educator, Winter 1992.

"Bringing meaning to math with a student-run store," by Deborah Black, Teachers' Lab, November/December 1992.

"Employer expectations for school mathematics," by Henry O. Pollak, Teachers' Lab, October 1992.

"Evaluating Problem Solving in Mathematics," by Walter Szetela and Cynthia Nicol, Educational Leadership, May 1992.

SECTION I: Overview of Assessment

Reprinted with permission from Stanford Law and Policy Review, Winter 1992-93.

Standards, Assessment, and Educational Quality

Lauren B. Resnick

Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15260

1993

Stanford Law and Policy Review, 4, 53-59.

Standards, Assessment, and Educational Quality

by

Lauren B. Resnick

National performance standards can play a vital role in systemic education reform.

American schools are underachieving institutions. Their aspirations are too low and they are not working up to capacity. The reason lies not in students learning less than before. Indeed, scores on tests of "basic skills" have risen over the past two decades.[1] But students are not learning the skills and knowledge they will need in the future. Why is this so?

Schooling as we know it today was designed early in this century.[2] It aimed to educate a small elite of future leaders— managers, engineers, physicians, lawyers, and other professionals— to use their minds well. For the majority of students, however, educational aims were much more modest. The goal was to teach basic citizenship and to inculcate the limited skills young people would need to take their places as workers in an economy needing many more willing hands than active minds. It was a mass production form of education suited to the mass production economy of the time.[3] Broad, liberating education for the many was considered unnecessary and unachievable. American tests, the standardized tests that report on children's and schools' "grade level", are the products of mass production education.[4] They assess students' command of disembodied bits of information, not their ability to analyze complex situations or marshal knowledge to solve problems.[5] Yet these tests substantially control what is taught in schools.

The strategy seemed to work for many decades: Living standards rose for more and more Americans and American democracy seemed secure.[6] But the world at the beginning of the next century will bear few rewards for individuals or nations that limit themselves to educating

Dr. Lauren B. Resnick is the director of the Learning Research and Development Center at the University of Pittsburgh and a Professor in the Department of Psychology and School of Education at the Univ. of Pittsburgh. She is also the co-director of the New Standards Project, which is developing a new national student performance assessment system. *Preparation of this paper was supported by grants from the Pew Charitable Trusts and the John D. and Catherine T. MacArthur Foundation for the work of the New Standards Project. An earlier version of the paper was presented at the annual meeting of the American Educational Research Association, April 1992, as part of the symposium: "The Federal Reform of Education: Boon or Bane for American Public Schools."*

only a few to think.[7] To maintain a high-wage economy, almost all individuals will have to think their way through their workdays—analyzing problems, proposing solutions, troubleshooting and repairing equipment, communicating with others, and managing resources of time and materials. For the first time since the Industrial Revolution, the human resource needs of a vibrant economy and the civic requirements of a truly participatory democracy are converging. The time has come for American schools to set their sights higher, to move from their inherited preoccupation with low-level fact and skill learning to the goals of thinking, reasoning, and problem-solving for every student.

Doing so will require the most thorough revision of aspiration and practice that any set of institutions has ever known. It will necessitate setting new standards of quality and mobilizing all resources to ensure that every single child, regardless of the child's race, language, origin, or presumed native ability, has the learning opportunities needed to meet those standards. It will mean re-educating teachers to work in new ways and giving them the authority they will need to set a new course with their students. New methods of assessing student achievement will be essential in this transformation.

> The time has come for American schools to set their sights higher and address the goals of thinking, reasoning and problem-solving.

New forms of assessment, assessments based on complex task performances, scored by trained and thoughtful judges, can release educators from the grip of testing programs that drive instructional attention away from thoughtfulness and complex applications of knowledge. They can exemplify new standards for achievement and provide clear representations of what students should now be striving to learn and teachers teach.

Standards and assessment alone, however, will work no magic. Real educational improvement requires interlocking and coherent changes in several components of the education system. These include curriculum, textbooks, teacher preparation, and continuing professional development. Further, for these elements to sustain themselves, there will need to be fundamental changes in the way schools are managed and in how schools relate to families, communities, and the social service delivery system. Stimulating and enabling these interlocking changes constitutes a *systemic reform* policy—the only kind of policy likely to produce the new levels of student achievement that are sought.

PERFORMANCE STANDARDS AND SYSTEMIC CHANGE

Implementing systemic change requires attention not only to how the elements of a system function together in equilibrium, but to how they might influence one another in a period of disequilibrium, and which elements might be most susceptible to organized, intentional modification. In some countries, an effort at national improvement of education would begin with attention to curriculum and, perhaps, organizational structure, including teacher certification requirements.[8] Both of these would fall under the purview of a ministry of education with the power, after consultation appropriate to the country's processes, to impose the curriculum and the new forms of organization. Textbooks, exams, and the content of teacher education programs would change in due course in response to the mandated curriculum.

In countries that, like us, have traditions of local rather than national control of education, this centralized approach to changing the system is not available. Some countries, most notably Britain, have responded to this condition by overthrowing significant aspects of the local control tradition and moving toward a national curriculum.[9] We in the United States are trying to induce systemic changes without federalizing education and without creating a controlling national curriculum. We need a different point of departure. What are the possibilities?

Three frequently proposed starting points for reform are textbooks, teacher education, and education governance. All are crucial to a full program of systemic reform. But none of the three serves as a promising starting point. Consider textbooks first. These serve as a kind of *de facto* national curriculum, and changes in textbooks would strongly influence what is taught and learned. But because textbook publishers respond primarily to market incentives, a radical change in demand from purchasers (i.e., educators) must be created. Textbook publishing can be expected to follow, not lead, the reform effort.

The prospects are not much better for teacher education as a starting point. The argument for reforming teacher education is compelling; only a quite differently

prepared teaching force will be able to educate students in the new, more demanding ways that are required for the future. But teacher education in this country is largely controlled by institutions with even greater traditions of "local" control than the public schools. Individual faculty at colleges and universities substantially control their own programs of instruction. No one has proposed a convincing and, at least for the moment,

> We in the United States are trying to induce systemic change without federalizing education and creating a controlling national curriculum.

politically viable means of directly inducing the hundreds of institutions responsible for educating teachers to make radical changes in what they do themselves or demand of future teachers. Furthermore, even with major changes in pre-service teacher education, it would take a long time to change schooling practice. Newcomers to any profession are not well positioned to take the lead in changing practice. For many years to come, the vast majority of teachers in place will have been educated in the "old" ways. Means must be found to more directly affect the practices of those already in the teaching force.

What about changing the management structure of education—giving local educators and parents more direct decision-making power? There seems little doubt, for reasons discussed below, that such changes are essential to overall systemic reform. Apart from a few schools with substantial outside resources, however, changes in management and decision making have, so far, rarely produced significant changes in curriculum, teaching, and learning. Those that have reached such results have almost always depended upon exceptional individual leaders.[10] These leaders often negotiate special privileges for the school—for example, freedom from certain mandated tests during an experimental period, the right to choose faculty in ways that are not standard in their district—and sometimes raise extra funds from foundations and other donors. Such schools are tolerated as exceptional experiments that do not really challenge the normal ways in which the system works. Without a change in the surrounding system, exceptional schools of this kind are fragile and likely to be driven back to ordinary ways of proceeding as soon as they become too visibly successful or their special advocates move on to new challenges.

Performance standards are a promising start for systemic change. We need an education system in which good schools that have high expectations for students and that work hard to meet them can thrive—where good schools are the norm rather than exceptions needing special protection. We should organize education so that it does not depend upon a few exceptional people putting out extraordinary effort, but relies instead upon thousands of competent and committed people working in concert with, rather than against, the "system." Performance standards represent the best means to achieve this. Such standards provide tangible goals that students can strive to achieve. Standards also allow teachers to measure student performance against an objective criterion of excellence. If we can agree on national standards for student achievement and create conditions in school systems all over the country in which those standards are internalized and made the centerpiece of educators' and students' efforts, there is a good probability that curriculum, professional development, textbooks, and, eventually, teacher preparation can be changed so that the entire system is working toward the standards.

INTERNALIZED STANDARDS: KEY TO A NEW CULTURE OF TEACHING

This approach may sound like a prescription for federal, or at least national, tests or exams tied to standards set by a committee of experts. It is not. National standards alone will not achieve our goals. Standards must be *internalized*. Unless standards are held as personal goals—by teachers first, and eventually by students—little in the way of profound educational change can result from a standard-setting and assessment process. Let us consider why this is so.

If a national test embodying the highest standards of a thinking curriculum could be created in Washington and given to every student in America, with test scores sent back to schools from a national office, educators might hope to change their school's performance—especially if some consequences for themselves or their students were attached to test scores. But if educators did not understand in a profound manner the differences between high and low scores, they would not know how to direct their efforts. Additionally, if educators were not personally invested in the standards and assessments based on them, the various forms of cheating and "working around the system" that have been amply documented in

studies of test-based, high-stakes reform efforts would be likely to dominate educators' response.[11]

For these reasons, an effective role for the federal government—or any national agency—in educational reform cannot be the apparently simple one of imposing standards from above or afar. The real task is not to announce new standards, but to create an educational culture in which working to achieve those standards pervades the system because the standards have become the internalized goals of everyone in the system.

Leading businesses working to become high-performance organizations understand this point.[12] They do not attempt to raise quality by putting in tougher reject standards at the end of the production line. Instead, they mount a complex process of engaging employees to understand new business requirements, reorganize their work patterns, and set the quality standards toward which they will work. A new culture, not a set of mechanical procedures, is at the heart of a transformation in quality.

> Unless standards are held as personal goals by teachers and students, little in the way of profound educational change can result.

Education can do no less. Changing standards and assessments by themselves will not substantially change real learning. Those changes will come only when teachers act as professionals who internalize and voluntarily work toward the standards of their field. Promoting that internalization must lie at the core of any strategy for educational reform.

DEVELOPING INTERNALIZED STANDARDS: SOME PRINCIPLES FOR ACTION

How can we achieve this goal? We need some principles for designing a standards-based systemic reform effort. The principles that follow guide the work of the New Standards Project —a partnership of seventeen states and six major school districts which have joined together to develop alternative approaches to setting education standards and assessing student achievement.[13] They are working to develop a *system* of assessments in which varied local assessments can be benchmarked to a shared set of national standards.

Engage Teachers as Central Actors in Standard Setting and Assessment Development

This approach is not limited to a few teachers who represent the teacher point of view in a committee of experts, but rather every teacher whose students will be judged by the standards and assessment. Teachers in schools throughout the country must be involved in developing assessment tasks and scoring student work. The primary actors in the work of the New Standards Project have been teachers. Teachers have been our initial assessment task developers. Nearly 500 teachers participated in administering performance assessments in the New Standards fourth grade trial assessments in mathematics and literacy last spring. This same group, led by a set of specially trained lead teachers, will also be responsible for scoring the assessments.[14]

Create a Culture Incorporating Professional Discussion of Standards for Student Work

This guideline represents an extension of the previous principle, specifying *how* the engagement of teachers in the standards process ought to proceed. Imagine that all the teachers involved in scoring assessments participated in extensive professional discussion of why particular student performances warranted particular scores. Teachers working in this kind of a standards and assessment system would become self-conscious, reflective judges of student work. Further, if they participated from time to time not just in applying scoring criteria developed by others but in developing criteria and arguing for them, the standards would come to be "owned" by the teachers. Such standards would be internalized. Teachers would be motivated not just to raise scores, but to meet quality standards they understood, believed in, and had some degree of control over. In the New Standards Project, teachers administering the new assessments participated in orientation discussions in which curriculum was discussed and responses of students to the assessment tasks were reviewed. As the project unfolds, new participating teachers will become members of ongoing work groups in which they generate and critique tasks, analyze student performances, and design and apply scoring criteria.

Embed Most Assessment in Regular Curriculum Work Rather Than in "One Shot" Examinations

Assessments need not be limited to what students can do in scheduled, supervised exams. Experiments in many individual schools, and now in two entire states (Vermont and Kentucky) show that collected portfolios of students' work, much like the portfolios artists prepare for art juries, can provide rich evaluations of students' capabilities.[15] These portfolios might contain the results of extended projects, exhibitions prepared by the student, and special performances, thus applying to the academic realm the practices of the visual and performing arts, sports, and scouting "merit badges." In a portfolio system, assessments are not separate from the curriculum; they are part of it, a special window on the regular work of the student. The New Standards assessment system will place portfolios of individual student work at its heart. Scheduled performance exams will be used primarily to anchor and audit portfolio-based scores.

There are many reasons to use such curriculum-embedded assessments as the major methods to judge student performance. Not least, the extended activities that can be included in portfolios come closer to reflecting deep educational goals than do scheduled exam questions. Such assessments also give students multiple and varied opportunities to show their competencies. But one fundamental benefit is that curriculum-embedded assessments, by definition, can permeate the teaching and learning process. In the most effective uses of portfolio assessment, students, working with their teachers, play a major role in selecting what should be presented as their representative and best work. Perhaps the most important benefit of such a portfolio process involves its capacity to develop self-reflective judgment capacities on the part of both students and teachers. This is the kind of internalization process for which we are aiming.

Make Standard Setting a Public Process

Teachers and other educators must be central actors in developing education standards, but they cannot be solely responsible. Education is a social good, in which many elements of society have a stake. Unless there is a general consensus about educational goals and standards, educators will be subject to continual pressures that will weaken or disable their efforts to meet standards they may have adopted. In many European countries, after exams are given, examination questions are published and discussed on television and radio shows. We need an American version of this kind of public participation, one that can engage all segments of our varied communities—parents, employers, community and child advocates—in active discussion of education goals and standards. In the New Standards Project, we will be developing ways to use print and broadcast media to create opportunities for citizen groups to discuss examples of students' work on performance exams and curriculum-embedded assessments. Focus groups we have already run in several locations around the country show that people of many different backgrounds and levels of education are able to engage productively in such discussions.[16] Our task now is to work with states and localities to make opportunities for such discussions broadly available and to develop procedures for systematically using their opinions as part of a formal standard-setting process.

Without a general consensus about educational goals and standards, educators will be subject to pressures that will weaken or disable their efforts.

APPROPRIATE GOVERNMENT ROLES

Given this view of the role of standards and assessments in educational reform, what roles are appropriate for federal, state, and local agencies? Clearly, building central national tests intended for administration to all children in the nation would not be productive. Internalized standards, not externally imposed ones, are the only likely means of creating broad changes in educational effort. Even making administration of a national test voluntary would not help much. It is unlikely that individual teachers, schools or communities would be given an opportunity to volunteer, and they would certainly have little chance to directly influence assessment content or process. Thus, the effect of a voluntary, but centrally developed, single national test would not be very different from an imposed one. Assessments, like the specifics of curricula to which they are tied, need to engage the energies of those closest to instruction. Thus, assessments will need to include substantial local variation.

At the same time, broadly specified national curriculum standards, such as those that now exist for mathematics, can play a vital role in systemic education reform. In state after state today, and in many school districts as well, educator and citizen groups are grappling with the problem of setting goals and standards for their education systems. The need for well thought-out frameworks for specific reform is clear: In the New Standards Project, for example, development of mathematics assessments has proceeded more smoothly than literacy assessments, primarily because the standards created by the National Council of Teachers of Mathematics (NCTM) provided a common framework from the start.[17] At the same time, the NCTM standards provide plenty of room for localities—whether states or individual districts—to craft curricula and assessments suitable for their own communities. As an example, one NCTM standard used in grades K-4 for the mastery of geometry is the ability to describe, model, draw, and classify shapes.[18] Few have the resources to engage in the kind of standards development work that NCTM has done at a national level. Thus, an important role for the federal government involves support of national content standards development for other parts of the curriculum. Fortunately, the U.S. Department of Education has begun to fund development of national content standards, and we can now look forward to having such standards for most of the major school subjects.

It is states and localities—not the federal government—that will have the primary responsibility for putting programs of systemic reform into place.

It is states and localities, not the federal government, that will in the end have the primary responsibility for putting programs of systemic reform, including standards and assessments, into place. Traditionally, education has largely been a power reserved to the states.[19] Efforts by the federal government to directly impose curricula could be subject to legal challenge. But to say that the power to impose standards resides exclusively at the state level is to ignore the substantial American tradition of much more local control of education. Furthermore, state-imposed assessments would not be automatically more conducive to internalization of standards by educators and communities than national ones. To convince ourselves of this, we need only recall that it is state, not national, tests that have been used in the mandated assessment programs in which test scores have been allegedly manipulated.[20] To reap the benefits of internalized standards, states will need to leave much discretion in the matter of assessment to local education authorities and even individual schools. A portfolio system, in which some assessment tasks are common to large numbers of students and others are locally selected by schools and individual teachers, can provide the necessary balance between local initiative and statewide standards.

Localities want curricula, and thus assessments, crafted to their own needs. At the same time, they want to know that they are competitive beyond the confines of their local communities; few want to educate their children in ways that would not permit them to prosper beyond local borders. To achieve these objectives, methods must be found to *link* the varied assessments that will be developed by localities to shared state and national standards. Simple statistical formulas for converting scores on one test into equivalent scores on another are unlikely to work; to legitimately apply such formulas, the assessments would need to be so alike that it is questionable that local options would be practicable.[21] A number of proposals for alternative ways of linking assessments are under discussion. These include administering a national "reference exam" to a sample of students in a state or district, including common national "anchor tasks" in a state or district's assessments, and arranging for "cross-grading" of one jurisdiction's exams by another's trained scorers.[22] None of these, however, is appropriate if different assessments are not first established as "reasonable" tests of the same content and objectives for learning. This can only be determined through an organized process of social and professional judgment. For this reason, a standards certification board of some kind will almost certainly be needed as part of a national assessment system.[23]

CONCLUSION

The program outlined here is both ambitious and optimistic. It assumes that in the future American students will need to learn skills in new ways and at much higher standards than ever before. This approach suggests, in turn, that teachers will need to teach in new ways and with different expectations for their students than many have held in the past. Performance standards, represented in assessments, are a powerful means to begin the process

of systemic educational change. Such standards provide tangible goals that students can strive to achieve, allowing teachers to measure student performance against an objective criterion of excellence. Assessments can be tailored to local school environments yet linked with broader state and national standards. Extensive participation by teachers in designing and implementing new assessments can serve as part of a professional development program that will transform the new standards into internalized, personal goals of teachers and students throughout the country. Standards and assessments used in this way promise to prepare America to compete with brain rather than brawn in the next century. ▲

NOTES

[1] Linda Darling-Hammond, *The Implications of Testing Policy for Quality and Equality*, 73 PHI DELTA KAPPAN 220, 221 (1991).

[2] RAY MARSHALL & MARC TUCKER, THINKING FOR A LIVING: EDUCATION AND THE WEALTH OF NATIONS 13-27 (1992).

[3] *Id.*

[4] Darling-Hammond, *supra* note 1, at 221-222.

[5] Marshall & Tucker, *supra* note 2, at 13-27.

[6] *Id.* at 31.

[7] *Id.* at 64-69.

[8] COMMISSION ON THE SKILLS OF THE AMERICAN WORKFORCE, NATIONAL CTR. ON EDUC. AND THE ECONOMY, AMERICA'S CHOICE: HIGH SKILLS OR LOW WAGES! 94-102 (1990).

[9] Paul Maston, *Response on Poor Pupils Was 'Too Slow'*, THE DAILY TELEGRAPH, Dec. 16, 1991, at 2.

[10] Marshall & Tucker, *supra* note 2, at 109-110.

[11] *Cf.* Darling-Hammond, *supra* note 1, at 221.

[12] Marshall & Tucker, *supra* note 2, at 110.

[13] Members of the New Standards Project are: the states of Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Iowa, Kentucky, Maine, New York, Oregon, South Carolina, Texas, Vermont, Virginia, and Washington; and the districts of Forth Worth, Pittsburgh, Rochester, San Diego, White Plains, and New York City. Together they enroll close to 50% of all U.S. public school children. New Standards is a joint undertaking of the Learning Research and Development Center at the University of Pittsburgh and the National Center on Education and the Economy in Rochester, New York. The program is described in LAUREN B. RESNICK & MARC TUCKER, NATIONAL CTR. ON EDUC. AND THE ECONOMY, THE NEW STANDARDS PROJECT 1992-1995, A PROPOSAL (1992).

[14] *Id.* at 40.

[15] Darling-Hammond, *supra* note 1, at 224.

[16] VINCENT J. BRECLIO, NATIONAL CTR. ON EDUC. AND THE ECONOMY, PUBLIC RESPONSE TO THE NEW STANDARDS PROGRAM (1991).

[17] NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS, CURRICULUM AND EVALUATION STANDARDS FOR SCHOOL MATHEMATICS (1989).

[18] *Id.* at 48.

[19] Charles F. Faber, *Is Local Control of the Schools Still a Viable Option?*, 14 HARV. J.L. & PUB. POL. 447.

[20] Darling-Hammond, *supra* note 1, at 221.

[21] Robert L. Linn, Educational Assessment: Expanded Expectations and Challenges, Address at the American Psychological Association Annual Meeting (Aug. 16, 1992).

[22] ROBERT L. LINN, NATIONAL CTR. FOR RESEARCH ON EVALUATION, STANDARDS AND STUDENT TESTING, CROSS-STATE COMPARABILITY OF JUDGMENTS OF STUDENT WRITING: RESULTS FROM THE NEW STANDARDS PROJECT, 21-22 (1991).

[23] NATIONAL COUNCIL ON EDUC. STANDARDS AND TESTING, RAISING STANDARDS FOR AMERICAN EDUCATION 35-36 (1992).

BEST COPY AVAILABLE

Reprinted with permission from PHI DELTA KAPPAN, May 1989.

A True Test: Toward More Authentic and Equitable Assessment

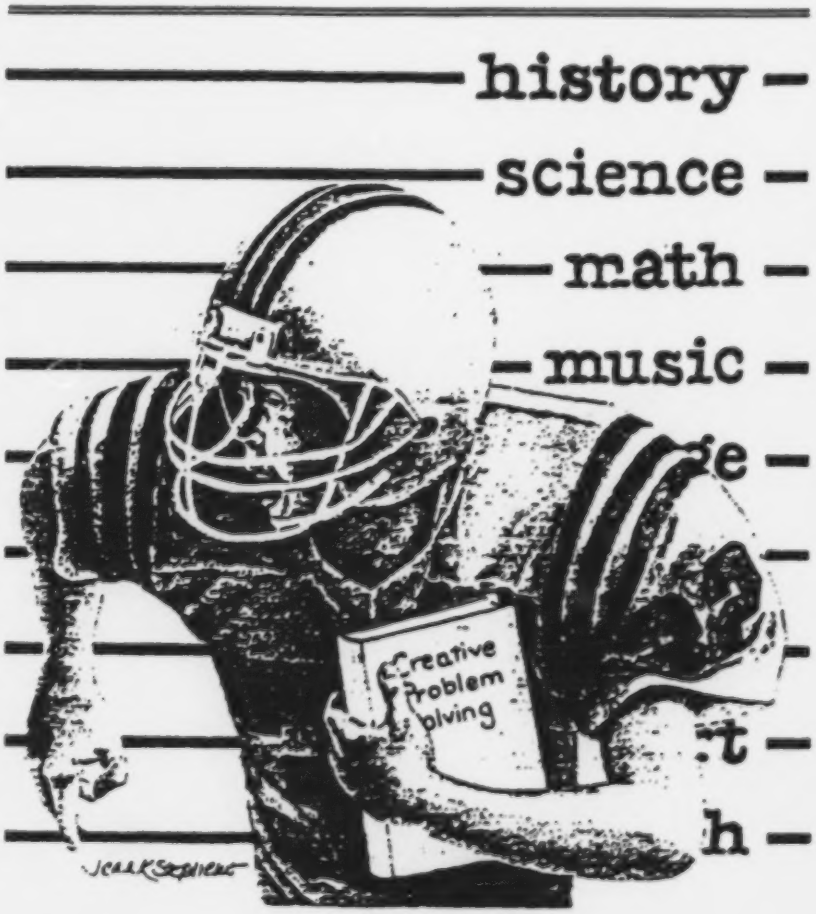
As long as we hold simplistic monitoring tests to be adequate models of and incentives for reaching national intellectual standards, Mr. Wiggins warns, student performance, teaching, and our thinking and discussion about assessment will remain flaccid and uninspired.

BY GRANT WIGGINS

WHEN AN educational problem persists despite the well-intentioned efforts of many people to solve it, it's a safe bet that the problem hasn't been properly framed. Assessment in education has clearly become such a problem, since every state reports above-average scores on norm-referenced achievement tests and since everyone agrees (paradoxically) that such tests shouldn't drive instruction but that their number and influence should nevertheless increase.[1] More ominously, we seem unable to see any moral harm in bypassing context-sensitive human judgments of human abilities in the name of statistical accuracy and economy.

We haven't moved beyond lamenting these problems, because we have failed to stop and ask some essential questions: Just what are tests meant to do? Whose purposes do they (and should they) serve? Are large-scale testing programs necessary? When are tests that are designed to monitor accountability harmful to the educational process? Need they be so intrusive? Is there an approach to upholding and examining a school's standards that might actually aid learning?

But we won't get far in answering these questions until we ask the most basic one: What *is* a true test? I propose a *radical* answer, in the sense of a return to the roots; we have lost sight of the fact that a true test of intellectual ability requires *the performance of exemplary tasks*. First, authentic assessments replicate the challenges and standards of performance that typically face writers, businesspeople, sci-



GRANT WIGGINS is a senior associate with the National Center on Education and the Economy, Rochester, N.Y., and a special consultant on assessment for the Coalition of Essential Schools.

entists, community leaders, designers, or historians. These include writing essays and reports, conducting individual and group research, designing proposals and mock-ups, assembling portfolios, and so on. Second, legitimate assessments are responsive to individual students and to school contexts. Evaluation is most accurate and equitable when it entails human judgment and dialogue, so that the person tested can ask for clarification of questions and explain his or her answers.

A genuine test of intellectual achievement doesn't merely check "standardized" work in a mechanical way. It reveals achievement on the essentials, even if they are not easily quantified. In other words, an authentic test not only reveals student achievement to the examiner, but also reveals to the test-taker the actual challenges and standards of the field.

To use a medical metaphor, our confusion over the uses of standardized tests is akin to mistaking pulse rate for the total effect of a healthful regimen. Standardized tests have no more effect on a student's intellectual health than taking a pulse has on a patient's physical health. If we want standardized tests to be authentic, to help students learn about themselves and about the subject matter or field being tested, they must become more than merely indicators of one superficial symptom.

Reform begins, then, by recognizing that the test is central to instruction. Any tests and final exams *inevitably* cast their shadows on all prior work. Thus they not only monitor standards, but also set them. Students acknowledge this truth with their plaintive query, Is this going to be on the test? And their instincts are correct; we should not feel despair about such a view. The test always sets the de facto standards of a school despite whatever else is proclaimed. A school *should* "teach to the test." The catch is that the test must offer students a genuine intellectual challenge, and teachers must be involved in designing the test if it is to be an effective point of leverage.

SETTING STANDARDS

We need to recognize from the outset that the problems we face are more ecological (i.e., political, structural, and economic) than technical. For example, Norman Frederiksen, a senior researcher with the Educational Testing Service (ETS), notes that "situational tests are not widely used in testing programs because of considerations having to do with cost and efficiency."[2] In order to overcome the resistance to using such situational tests, we must make a powerful case to the public (and to teachers habituated to short-answer tests as an adequate measure of ability) that a standardized test of intellectual ability is a contradiction in terms. We must show that influential "monitoring" tests are so irrelevant (and even harmful) to genuine intellectual standards that their cost – to student learning and teacher professionalism – is too high, however financially efficient they may be as a means of gathering data.

The inescapable dilemma presented by mass testing is that using authentic standards and tasks to judge intellectual ability is labor-intensive and time-consuming. Examiners must be trained, and multiple, contexnual tests of the students must be conducted. Genuine tests also make it more difficult to compare, rank, and sort because they rarely involve one simple, definitive test with an unambiguous result and a single residue number. Therefore, as long as tests are thought of only in terms of accountability, real reforms will be thwarted. After all, why do we need to devise more expensive tests if current data are reliable? When we factor in the self-interest of test companies and of colleges and school districts, we can see that resistance to reform is likely to be strong.

Using authentic standards and tasks to judge intellectual ability is labor-intensive and time-consuming.

The psychometricians and the accountants are not the villains, however. As I have noted elsewhere, teachers fail to understand their own unwitting role in the growth of standardized testing.[3] Mass assessment resulted from legitimate concern about the failure of the schools to set clear, justifiable, and consistent standards to which it would hold its graduates and teachers accountable. But the problem is still with us: high school transcripts tell us nothing about what a student can actually do. Grades and Carnegie units hide vast differences between courses and schools. An A in 11th-grade English may mean merely that a student was dutiful and able to fill in blanks on worksheets about juvenile novels. And it remains possible for a student to pass all of his or her courses and still remain functionally and culturally illiterate.

But the solution of imposing an efficient and "standard" test has an uglier his-

tory. The tests grew out of the "school-efficiency" movement in the years between 1911 and 1916, a time oddly similar to our own. The movement, spearheaded by the work of Franklin Bobbitt, was driven by crude and harmful analogies drawn from Frederick Taylor's management principles, which were used to improve factory production. Raymond Callahan notes that the reformers, then as now, were far too anxious to satisfy external critics and to reduce complex intellectual standards and teacher behaviors to simple numbers and traits.[4] Implicitly, there were signs of hereditarian and social-class-based views of intelligence; the tests were used as sorting mechanisms at least partly in response to the increased heterogeneity of the school population as a result of the influx of immigrants.[5]

The "standards" were usually cast in terms of the increased amount of work to be demanded of teachers and students. As George Strayer, head of the National Education Association (NEA) Committee on Tests and Standards for School Efficiency, reported, "We may not hope to achieve progress except as such measuring sticks are available." A school superintendent put it more bluntly: "The results of a few well-planned tests would carry more weight with the businessman and parent than all the psychology in the world."[6]

Even with unionization and the insights gained from better education, modern teachers still fall prey to the insistent claims of noneducation interests. The wishes of college admissions officers, of employers, of budget makers, of schedulers, and even of the secretaries who enter grades on computers often take precedence over the needs of students to be properly examined and the needs of teachers to deliberate and confer about effective test design and grading.

Thus, when teachers regard tests as something to be done as quickly as possible after "teaching" has ended in order to shake out a final grade, they succumb to the same flawed logic employed by the test companies (with far less statistical justification). Such acquiescence is possible only when the essential ideas and priorities in education are unclear or have been lost. If tests serve only as administrative monitors, then short-answer, "objective" tests – an ironic misnomer[7] – will suffice (particularly if one teaches 128 students and has only a single day in which to grade final exams). However, if a test is seen as the heart and soul of the educational enterprise, such reductionist shortcuts, such high student/teacher ratios, and such dysfunctional allocation of time and resources will be seen as intolerable.

Schools and teachers do *not* tolerate the same kind of thinking in athletics, the arts, and clubs. The requirements of the game, recital, play, debate, or science fair are clear, and those requirements determine the use of time, the assignment of personnel, and the allocation of money. Far more time – often one's spare time – is devoted to insuring adequate practice and success. Even in the poorest schools, the ratio of players to interscholastic coaches is about 12 to 1.[8] The test demands such dedication of time; coaching requires one-to-one interaction. And no one complains about teaching to the test in athletic competition.

We need to begin anew, from the premise that a testing program must address questions about the inevitable impact of tests (and scoring methods) on students and their learning. We must ask different questions. What kinds of challenges would be of most educational value to students? What kinds of challenges would give teachers useful information about the abilities of their students? How will the results of a test help students know their strengths and weaknesses on essential tasks? How can a school adequately communicate its standards to interested outsiders and justify them, so that standardized tests become less necessary and less influential?

AUTHENTIC TESTS

Tests should be central experiences in learning. The problems of administration, scoring, and between-school comparisons should come only after an authentic test had been devised – a reversal of the current practice of test design.

If we wish to design an authentic test, we must first decide what are the actual performances that we want students to be good at. We must design those performances first and worry about a fair and thorough method of grading them later. Do we judge our students to be deficient in writing, speaking, listening, artistic creation, finding and citing evidence, and problem solving? Then let the tests ask them to write, speak, listen, create, do original research, and solve problems. Only then need we worry about scoring the performances, training the judges, and adapting the school calendar to insure thorough analysis and useful feedback to students about results.

To design an authentic test, we must first decide what are the actual performances that we want students to be good at.

This reversal in thinking will make us pay more attention to what we mean by *evidence of knowing*. Mastery is more than producing verbal answers on cue; it involves thoughtful understanding, as well. And thoughtful understanding implies being able to do something effective, transformative, or novel with a problem or complex situation. An authentic test enables us to watch a learner pose, tackle, and solve slightly ambiguous problems. It allows us to watch a student marshal evidence, arrange arguments, and take purposeful action to address the problems.[9] Understanding is often best seen in the ability to *criticize* or extend knowledge, to explain and explore the limits and assumptions on which a theory rests. Knowledge is thus displayed as thoughtful know-how – a blend of good judgment, sound habits, responsiveness to the problem at hand, and control over the appropriate information and context. Indeed, genuine mastery usually involves even more: doing something with grace and style.

To prove that an answer was not an accident or a thoughtless (if correct) response, multiple and varied tests are required. In performance-based areas we do not assess competence on the basis of one performance. We repeatedly assess a student's work – through a portfolio or a season of games. Over time and in the context of numerous performances, we observe the *patterns* of success and failure and the reasons behind them. Traditional tests – as arbitrarily timed, superficial exercises (more like drills on the practice field than like a game) that

are given only once or twice – leave us with no way of gauging a student's ability to make progress over time.

We typically learn too much about a student's short-term recall and too little about what is most important: a student's habits of mind. In talking about *habits of mind*, I mean something more substantive than "process" skills divorced from context – the formalism decried by E. D. Hirsch and others. For example, a new concept – say, irony or the formula F = ma – can be learned as a habit or disposition of mind for effortlessly handling information that had previously been confusing.[10] As the word *habit* implies, if we are serious about having students display thoughtful control over ideas, a single performance is inadequate. We need to observe students' *repertoires*, not rote catechisms coughed up in response to pat questions.

The problem is more serious than it first appears. The difficulty of learning lies in the breaking of natural but dysfunctional habits. The often-strange quality of new knowledge can cause us to unwittingly misunderstand new ideas by assimilating them into our old conceptions; this is particularly true when instruction is only verbal. That is why so many students who do well on school tests seem so thoughtless and incompetent in solving real-world problems. For example, the research done at Johns Hopkins University demonstrates how precarious and illusory "knowledge" of physics really is, when even well-trained students habitually invoke erroneous but plausible ideas about force on certain problems.[11]

The true test is so central to instruction that it is known from the start and repeatedly taken *because* it is both central and complex – equivalent to the game to be played or the musical piece to be performed. The true test of ability is to perform consistently well tasks whose criteria for success are known and valued. By contrast, questions on standardized tests are usually kept "secure," hidden from students and teachers, and they thus contradict the most basic conditions required for learning.[12] (Of course, statistical validity and reliability *depend* on the test being secret, and, when a test is kept secret, the questions can be used again.)

Designing authentic tests should involve knowledge use that is forward-looking. We need to view tests as "assessments of enablement," to borrow Robert Glaser's term. Rather than merely judging whether students have learned what was taught, we should "assess knowledge in terms of its constructive use for further learning. . . . [We should assess reading ability] in a way that takes into account that the purpose of learning to read is to enable [students] to learn from reading."[13] All tests should involve students in the actual challenges, standards, and habits needed for success in the academic disciplines or in the workplace: conducting original research, analyzing the research of others in the service of one's research, arguing critically, and synthesizing divergent viewpoints. Within reasonable and reachable limits, a real test replicates the authentic intellectual challenges facing a person in the field. (Such tests are usually also the most engaging.)

> Most so-called "criterion-referenced" tests are inadequate because the problems are contrived, and the cues are artificial.

The practical problems of test design can best be overcome by thinking of academic tests as the intellectual equivalent of public "performances." To enable a student is to help him or her make progress in handling complex tasks. The novice athlete and the novice actor face the same challenges as the seasoned professional. But school tests make the complex simple by dividing it into isolated and simplistic chores – as if the student need not practice the true test of performance, the test of putting all the elements together. This apparently logical approach of breaking tasks down into their components leads to tests that assess only artificially isolated "outcomes" and provide no hope of stimulating genuine intellectual progress. As a result, teaching to such tests becomes mechanical, static, and disengaging. Coaches of musicians, actors, debaters, and athletes know better. They know that what one learns in drill is never adequate to produce mastery.

That is why most so-called "criterion-referenced" tests are inadequate: the problems are contrived, and the cues are artificial. Such tests remove what is central to intellectual competence: the use of judgment to recognize and pose complex problems as a prelude to using one's discrete knowledge to solve them. Authentic challenges – be they essays, original research, or artistic performances – are inherently ambiguous and open-ended. As Frederiksen has said:

> Most of the important problems one faces are ill-structured, as are all the really important social, political, and scientific problems. . . . But ill-structured problems are not found in standardized achievement tests. . . . Efficient tests tend to drive out less efficient tests, leaving many important abilities untested and unaught. . . . All this reveals a problem when we consider the influence of an accountability system in education. . . . We need a much broader conception of what a test is.[14]

Put simply, what the student needs is a test with more sophisticated criteria for judging performance. In a truly authentic and criterion-referenced education, far more time would be spent teaching and testing the student's ability to understand and internalize the criteria of genuine competence. What is so harmful about current teaching and testing is that they frequently reinforce – unwittingly – the lesson that mere right answers, put forth by going through the motions, are adequate signs of ability. Again, this is a mistake rarely made by coaches, who know that their hardest and most important job is to raise the standards and expectations of their students.

EXAMPLES OF AUTHENTIC TESTS

Let us examine some tests and criteria devised by teachers working to honor the ideas I've been discussing under the heading of "exhibition of mastery" – one of the nine "Common Principles" around which members of the Coalition of Essential Schools have organized their reform efforts.[15] Here are two examples of final exams that seem to replicate more accurately the challenges facing experts in the field.

An oral history project for ninth-grad-

ers.[16] You must complete an oral history based on interviews and written sources and present your findings orally in class. The choice of subject matter will be up to you. Some examples of possible topics include: your family, running a small business, substance abuse, a labor union, teenage parents, or recent immigrants. You are to create three workable hypotheses based on your preliminary investigations and come up with four questions you will ask to test each hypothesis.

To meet the criteria for evaluating the oral history project described above, you must:

- investigate three hypotheses;
- describe at least one change over time;
- demonstrate that you have done background research;
- interview four appropriate people as sources;
- prepare at least four questions related to each hypothesis;
- ask questions that are not leading or biased;
- ask follow-up questions when appropriate;
- note important differences between fact and opinion in answers that you receive;
- use evidence to support your choice of the best hypothesis; and
- organize your writing and your class presentation.

A course-ending simulation/exam in economics.[17] You are the chief executive officer of an established firm. Your firm has always captured a major share of the market, because of good use of technology, understanding of the natural laws of constraint, understanding of market systems, and the maintenance of a high standard for your product. However, in recent months your product has become part of a new trend in public tastes. Several new firms have entered the market and have captured part of your sales. Your product's proportional share of total aggregate demand is continuing to fall. When demand returns to normal, you will be controlling less of the market than before.

Your board of directors has given you less than a month to prepare a report that solves the problem in the short run and in the long run. In preparing the report, you should: 1) define the problem, 2) prepare data to illustrate the current situation, 3) prepare data to illustrate conditions one year in the future, 4) recommend action for today, 5) recommend action over the next year, and 6) discuss where your company will be in the market six months from today and one year from today.

The tasks that must be completed in the course of this project include:

- deriving formulas for supply, demand, elasticity, and equilibrium;
- preparing schedules for supply, demand, costs, and revenues;
- graphing all work;
- preparing a written evaluation of the current and future situation for the market in general and for your company in particular;
- preparing a written recommendation for your board of directors;
- showing aggregate demand today and predicting what it will be one year hence; and
- showing the demand for your firm's product today and predicting what it will be one year hence.

Connecticut has implemented a range of performance-based assessments in science, foreign languages, drafting, and small-engine repair, using experts in the field to help develop apt performance criteria and test protocols. Here is an excerpt from the Connecticut manual describing the performance criteria for foreign languages; these criteria have been derived from the guidelines of the American Council on the Teaching of Foreign Languages (ACTFL).[18] On the written test, students are asked to draft a letter to a pen pal. The four levels used for scoring are novice, intermediate, intermediate high, and advanced; they are differentiated as follows:

- *Novice.* Students use high-frequency words, memorized phrases, and formulaic sentences on familiar topics. Students show little or no creativity with the language beyond the memorized patterns.
- *Intermediate.* Students recombine the learned vocabulary and structures into simple sentences. Sentences are choppy, with frequent errors in grammar, vocabulary, and spelling. Sentences will be very simple at the low end of the intermediate range and will often read very much like a direct translation of English.
- *Intermediate high.* Students can write creative sentences, sometimes fairly complex ones, but not consistently. Structural forms reflecting time, tense, or aspect are attempted, but the result is not always successful. Student show an emerging ability to describe and narrate in paragraphs, but papers often read like academic exercises.
- *Advanced.* Students are able to join sentences in simple discourse and have sufficient writing vocabulary to express themselves simply, although the language may not be idiomatic. Students show good control of the most frequently used syntactic structures and a sense that they are comfortable with the target language and can go beyond the academic task.

Of course, using such an approach is time-consuming, but it is not impractical or inapplicable to all subject areas on a large scale. The MAP (Monitoring Achievement in Pittsburgh) testing program offers tests of critical thinking and writing that rely on essay questions and are specifically designed to provide diagnostic information to teachers and

students. Pittsburgh is also working, through its Syllabus-Driven Exam Program, to devise exemplary test items that are based more closely on the curriculum.[19]

On the state level, Vermont has recently announced that it will move toward a portfolio-based assessment in writing and mathematics, drawing on the work of the various affiliates of the National Writing Project and of the Assessment of Performance Unit (APU) in Great Britain. California has piloted performance-based tests in science and other subjects to go with its statewide essay-writing test.

RESPONSIVENESS AND EQUITY

Daniel Resnick and Lauren Resnick have proposed a different way of making many of these points. They have argued that American students are the "most tested" but the "least examined" youngsters in the world.[20] As their epigram suggests, we rarely honor the original meaning of the word *test*. Originally a *testum* was a porous cup for determining the purity of metal; later it came to stand for any procedures for determining the worth of a person's effort. To prove the value or ascertain the nature of a student's understanding implies that appearances can deceive. A correct answer can disguise thoughtless recall. A student might quickly correct an error or a slip that obscures thoughtful understanding; indeed, when a student's reasoning is heard, an error might not actually be an error at all.

The root of the word *assessment* reminds us that an assessor should "sit with" a learner in some sense to be sure that the student's answer *really* means what it seems to mean. Does a correct answer mask thoughtless recall? Does a wrong answer obscure thoughtful understanding? We can know for sure by asking further questions, by seeking explanation or substantiation, by requesting a self-assessment, or by soliciting the student's response to the assessment.

The problem can be cast in broader moral terms: the standardized test is disrespectful by design. Mass testing as we know it treats students as objects — as if their education and thought processes were similar and as if the reasons for their answers were irrelevant. Test-takers are not, therefore, treated as *human* subjects whose feedback is essential to the accuracy of the assessment. Pilot standardized tests catch many technical defects in test questions. However, responses to higher-order questions are inherently unpredictable.

TABLE 1.
An Item from the NAEP Science Test

Child's Name	Frisbee Toss (yds.)	Weight Lift (lbs.)	50-Yard Dash (secs.)
Joe	40	205	9.5
Jose	30	170	8.0
Kim	45	130	9.0
Sarah	28	120	7.6
Zabi	48	140	8.3

The standardized test is thus inherently inequitable. I am using the word *equity* in its original, philosophical meaning, as it is incorporated into the British and American legal systems. The concept is commonsensical but profound: blank laws and policies (or standardized tests) are inherently unable to encompass the inevitable idiosyncratic cases for which we ought always to make exceptions to the rule. Aristotle put it best: "The equitable is a correction of the law where it is defective owing to its universality."[21]

In the context of testing, equity requires us to insure that human judgment is not overrun or made obsolete by an efficient, mechanical scoring system. Externally designed and externally mandated tests are dangerously immune to the possibility that a student might legitimately need to have a question rephrased or might deserve the opportunity to defend an unexpected or "incorrect" answer, even when the test questions are well-structured and the answers are multiple choice. How many times do teachers, parents, or employers have to alter an evaluation after having an answer or action explained? Sometimes, students need only a hint or a slight rephrasing to recall and use what they know. We rely on human judges in law and in athletics because complex judgments cannot be reduced to rules if they are to be truly equitable. To gauge understanding, we must explore a student's answer; there must be some possibility of dialogue between the assessor and the assessed to insure that the student is fully examined.

This concern for equity and dialogue is not idle, romantic, or esoteric. Consider the following example from the National Assessment of Educational Progress (NAEP) science test, Learning by Doing, which was piloted a few years ago.[22] On one of the tasks, students were given three sets of statistics that supposedly derived from a mini-Olympics that some children had staged (see Table 1). The introductory text noted that the children "decided to make each event of the same importance." No other information that bears on the question was provided. The test presented the students with the results of three events from the competition.

The first question asked, Who would be the all-around winner? The scoring manual gives these instructions:

> Score 4 points for accurate ranking of the children's performance on each event and citing Zabi as the overall winner. Score 3 points for using a ranking approach . . . but misinterpreting performance on the dash event . . . and therefore, citing the wrong winner. Score 2 points for a response which cites an overall winner or a tie with an explanation that demonstrates some recognition that a quantitative means of comparison is needed. Score 1 point if the student makes a selection of an overall winner with an irrelevant or non-quantitative account or without providing an explanation. Score 0 for no response.

Makes sense, right? But now ask yourself how, using the given criteria, you would score the following response given by a third-grader:

> A. Who would be the all-around winner?
> No one.
> B. Explain how you decided who would be the all-around winner. Be sure to show your work.
> No one is the all-around winner.

The NAEP scorer gave the answer a score of 1. Given the criteria, we can see why. The student failed to give an explanation or any numerical calculations to support the answer.

But could that answer somehow be apt in the mind of the student? Could it be that the 9-year-old deliberately

and correctly answered "no one," since "all-around" could mean "winner of all events"? If looked at in this way, couldn't it be that the child was *more* thoughtful than most by deliberately *not* taking the bait of part B (which presumably would have caused the child to pause and consider his or her answer). The full sentence answer in part B – remember, this is a 9-year-old – is revealing to me. It is more emphatic than the answer to part A, as if to say, "Your question suggests I *should* have found one all-around winner, but I won't be fooled. I stick to my answer that no one was the all-around winner." (Note, by the way, that in the scorer's manual the word *all-around* has been changed to *overall*.) The student did not, of course, explain the answer, but it is conceivable that the instruction was confusing, given that there was no "work" needed to determine that "no one" was the all-around winner. One quick follow-up question could have settled the matter.

A moral question with intellectual ramifications is at issue here: Who is responsible for insuring that an answer has been fully explored or understood, the tester or the student? One reason to safeguard the teacher's role as primary assessor is that the most accurate and equitable evaluation depends on relationships that have developed over time between examiner and student. The teacher is the only one who knows what the student can or cannot do consistently, and the teacher can always follow up on confusing, glib, or ambiguous answers.

In this country we have been so enamored of efficient testing that we have overlooked feasible in-class alternatives to such impersonal testing, which are already in use around the world. The German *abitur* (containing essay and oral questions) is designed and scored by classroom teachers, who submit two possible tests to a state board for approval. The APU in Great Britain has for more than a decade developed tests that are designed for classroom use and that involve interaction between assessor and student.

What is so striking about many of the APU test protocols is that the assessor is meant to probe, prompt, and even teach, if necessary, to be sure of the student's actual ability and to enable the learner to learn from the assessment. In many of these tests the first answer (or lack of one) is not deemed a sufficient insight into the student's knowledge.[23] Consider, for example, the following sections from the assessor's manual for a mathematics test for British 15-year-olds covering the ideas of perimeter, area, and circumference.

1. Ask: "What is the perimeter of a rectangle?" [Write student answer.]
2. Present sheet with rectangle ABCD. Ask: "Could you show me the perimeter of this rectangle?" *If necessary, teach.*
3. Ask: "How would you measure the perimeter of the rectangle?" *If necessary, prompt for full procedure. If necessary, teach. . . .*
10. "Estimate the length of the circumference of this circle."
11. Ask: "What would you do to check your estimate?" [String is on the table.] *If no response, prompt for string.*
13. Ask: "Is there any other method?" *If student does not suggest using $C = \pi d$, prompt with,* "Would it help to measure the diameter of the circle?"

Who is responsible for insuring that an answer has been fully explored or understood, the tester or the student?

The scoring system works as follows: 1) unaided success; 2) success following one prompt from the tester; 3) success following a series of prompts; 4) teaching by the tester, prompts unsuccessful; 5) an unsuccessful response, and tester did not prompt or teach; 6) an unsuccessful response despite prompting and teaching; 7) question not given; and 8) unaided success where student corrected an unsuccessful attempt without help. The "successful" responses were combined into two larger categories called "unaided success" and "aided success," with percentages given for each.[24]

The Australians for years have used similar tasks and similarly trained teachers to conduct district- and statewide assessments in academic subject areas (much as we do in this country with the Advanced Placement exams). Teachers give tests made up of questions drawn from banks of agreed-upon items and then mark them. Reliability is achieved through a process called "moderation," in which teachers of the same subjects gather to compare results and to set criteria for grading.

To insure that professionalization is aided, not undermined, by national testing, the process of "group moderation" has been made a central feature of the proposed new national assessment system in Great Britain. The tests will be both teacher-given and standardized. But what is so admirable – and equitable – is that

> We must overcome the lazy habit of grading and scoring "on the curve" as a cheap way of setting and upholding standards.

the process of group moderation requires collective judgments about any discrepancies between grade patterns in different schools and between results in a given school and on the nationally standardized criterion-referenced test. Significantly, the process of moderation can, on occasion, override the results of the nationally standardized test:

> A first task of a moderation group would be to examine how well the patterns of the two matched for each group of pupils [comparing percentages of students assigned to each level]. . . . The meeting could then go on to explore discrepancies in the pattern of particular schools or groups, using samples of pupils' work *and knowledge of the circumstances of schools*. The group moderation would first explore any general lack of matching between the overall teacher rating distribution and the overall distribution of results on the national tests. The general aim would be to adjust the overall teacher rating results to match the overall results of the national tests; *if the group were to have clear and agreed reasons for not doing this, these should be reported . . . [and] departures could be approved if the group as a whole could be convinced that they were justified in particular cases.*[25] (Emphasis added)

At the school-site level in the U.S., we might consider the need for an oversight process akin to group moderation to insure that students are not subject to eccentric testing and grading – a committee on testing standards, for example. In short, what group moderation can provide is the kind of on-going professional development that teachers need and desire. Both equity in testing and reform of schooling ultimately depend on a more open and consensual process of establishing and upholding schoolwide standards.

A number of reasons are often cited for retaining "objective" tests (the design of which is usually quite "subjective"), among them: the unreliability of teacher-created tests and the subjectivity of human judgment. However, reliability is only a problem when judges operate in private and without shared criteria. In fact, multiple judges, when properly trained to assess actual student performance using agreed-upon criteria, display a high degree of inter-rater reliability. In the Connecticut foreign language test described above, on the thousands of student tests given, two judges using a four-point scoring system agreed on a student's score 85% of the time.[26] Criticisms of Advanced Placement exams that contain essay questions usually focus on the cost of scoring, not on problems of inter-rater reliability. Inadequate testing technology is a red herring. The real problem standing in the way of developing more authentic assessment with collaborative standard-setting is the lack of will to invest the necessary time and money.

True criterion-referenced tests and diploma requirements, though difficult to frame in performance standards, are essential for establishing an effective and just education system. We must overcome the lazy habit of grading and scoring "on the curve" as a cheap way of setting and upholding standards. Such a practice is unrelated to any agreed-upon intellectual standards and can reveal only where students stand in relation to one another. It tells us nothing about where they ought to be. Moreover, students are left with only a letter or number – with nothing to learn from.

Consider, too, that the bell-shaped curve is an *intended* result in designing a means of scoring a test, not some coincidental statistical result of a mass testing. Norm-referenced tests, be they locally or nationally normed, operate under the assumption that teachers have no effect – or only a random effect – on students.

> There is nothing sacred about the normal curve. It is the distribution most appropriate to chance and random activity. Education is a purposeful activity, and we seek to have the students learn what we have to teach. . . . [W]e may even insist that our efforts are unsuccessful to the extent that the distribution of achievement approximates the normal distribution.[27]

In addition, such scoring insures that, *by design*, at least half of the student population is always made to feel inept and discouraged about their work, while the other half often has a feeling of achievement that is illusory.

Grading on a curve in the classroom is even less justifiable. There is no statistical validity to the practice, and it allows teachers to continually bypass the harder but more fruitful work of setting and teaching performance criteria from which better learning would follow.

To let students show off what they

know and are able to do is a very different business from the fatalism induced by counting errors on contrived questions. Since standardized tests are designed to highlight differences, they often end up exaggerating them (e.g., by throwing out pilot questions that everyone answers correctly in order to gain a useful "spread" of scores).[28] And since the tasks are designed around hidden and often arbitrary questions, we should not be surprised if the test results end up too dependent on the native language ability or cultural background of the students, instead of on the fruit of their best efforts.

Tracking is the inevitable result of grading on a curve and thinking of standards only in terms of drawing exaggerated comparisons between students. Schools end up institutionalizing these differences, and, as the very word *track* implies, the standards for different tracks never converge. Students in the lower tracks are not taught and assessed in such a way that they become *better* enabled to close the gap between their current competence and ideal standards of performance.[29] Tracking simply enables students in the lower tracks to get higher grades.

In the performance areas, by contrast, high standards and the incentives for students are clear.[30] Musicians and athletes have expert performers constantly before them from which to learn. We set up different weight classes for wrestling competition, different rating classes for chess tournaments, and separate varsity and junior varsity athletic teams to nurture students' confidence as they slowly grow and develop their skills. We assume that progress toward higher levels is not only possible but is aided by such groupings.

The tangible sense of efficacy (aided by the desire to do well publicly and the power of positive peer pressure) that these extracurricular activities provide is a powerful incentive. Notice how often some students will try to sneak back into school after cutting class to submit themselves to the rigors of athletics, debate, or band practice – even when they are not the stars or when their team has an abysmal record.[31]

CRITERIA OF AUTHENTICITY

From the arguments and examples above, let me move to a consideration of a set of criteria by which we might distinguish authentic from inauthentic forms of testing.[32]

Structure and logistics. Authentic tests are more appropriately public, involving an actual audience, client, panel, and so on. The evaluation is typically based on judgment that involves multiple criteria (and sometimes multiple judges), and the judging is made reliable by agreed-upon standards and prior training.

Authentic tests do not rely on unrealistic and arbitrary time constraints, nor do they rely on secret questions or tasks. They tend to be like portfolios or a full season's schedule of games, and they emphasize student progress toward mastery.

Authentic tests require some collaboration with others. Most professional challenges faced by adults involve the capacity to balance individual and group achievement. Authentic tests recur, and they are worth practicing, rehearsing, and retaking. We become better educated by taking the test over and over. Feedback to students is central, and so authentic tests are more intimately connected with the aims, structures, schedules, and policies of schooling.

Intellectual design features. Authentic tests are not needlessly intrusive, arbitrary, or contrived merely for the sake of shaking out a single score or grade. Instead, they are "enabling" – constructed to point the student toward more sophisticated and effective ways to use knowledge. The characteristics of competent performance by which we might sort nonenabling from enabling tests might include: "The coherence of [the student's] knowledge, principled [as opposed to merely algorithmic] problem solving, usable knowledge, attention-free and efficient performance, and self-regulatory skills."[33]

Authentic tests are contextualized, complex intellectual challenges, not fragmented and static bits or tasks. They culminate in the student's own research or product, for which "content" is to be mastered as a *means*, not as an end. Authentic tests assess student habits and repertoires; they are not simply restricted to recall and do not reflect lucky or unlucky one-shot responses. The portfolio is the appropriate model: the general task is to assess longitudinal control over the essentials.[34]

Authentic tests are representative challenges within a given discipline. They are designed to emphasize realistic (but fair) complexity; they stress *depth* more than breadth. In doing so, they must necessarily involve somewhat ambiguous, ill-structured tasks or problems, and so they make student judgment central in posing, clarifying, and tackling problems.

Standards of grading and scoring. Authentic tests measure essentials, not easily counted (but relatively unimportant) errors. Thus the criteria for scoring them must be equally complex, as in the cases of the primary-trait scoring of essays or the scoring of ACTFL tests of foreign languages. Nor can authentic tests be scored on a curve. They must be scored with reference to authentic standards of performance, which students must understand to be inherent to successful performance.

Authentic tests are contextualized, complex intellectual challenges, not fragmented and static bits or tasks.

Authentic tests use multifaceted scoring systems instead of a single aggregate grade. The many variables of complex performance are disaggregated in judging. Moreover, self-assessment becomes more central.[35]

Authentic tests exist in harmony with schoolwide aims; they embody standards to which everyone in the school can aspire. This implies the need for schoolwide policy-making bodies (other than academic departments) that cross disciplinary boundaries and safeguard the essential aims of the school. At Alverno College in Milwaukee, all faculty members are both members of disciplinary departments and of "competency groups" that span all departments.

Fairness and equity. Rather than rely on right/wrong answers, unfair "distractors," and other statistical artifices to widen the spread of scores, authentic tests ferret out and identify (perhaps hidden) strengths. The aim is to enable the students to show off what they can do. Au-

thentic tests strike a constantly examined balance between honoring achievement, progress, native language skill, and prior fortunate training. In doing so, they can better reflect our intellectual values.

Authentic tests minimize needless, unfair, and demoralizing comparisons and do away with fatalistic thinking about results. They also allow appropriate room to accommodate students' learning styles, aptitudes, and interests. There is room for the quiet "techie" and the show-off prima donna in plays; there is room for the slow, heavy lineman and for the small, fleet pass receiver in football. In professional work, too, there is room for choice and style in tasks, topics, and methodologies. Why must all students be tested in the same way and at the same time? Why should speed of recall be so well-rewarded and slow answering be so heavily penalized in conventional testing?[36]

Authentic tests can be – indeed, should be – attempted by all students, with the tests "scaffolded up," not "dumbed down" as necessary to compensate for poor skill, inexperience, or weak training. Those who use authentic tests should welcome student input and feedback. The model here is the oral exam for graduate students, insuring that the student is given ample opportunity to explain his or her work and respond to criticism as integral parts of the assessment.

In authentic testing, typical procedures of test design are reversed, and accountability serves student learning. A model task is first specified. Then a fair and incentive-building plan for scoring is devised. Only then would reliability be considered. (Far greater attention is paid throughout to the test's "face" and "ecological" validity.)

> Only a humane and intellectually valid approach to evaluation can help us insure progress toward national intellectual fitness.

As I said at the outset, we need a new philosophy of assessment in this country that never loses sight of the student. To build such an assessment, we need to return to the roots of authentic assessment, the assessment of *performance of exemplary tasks*. We might start by adopting the manifesto in the introduction of the new national assessment report in Great Britain, a plan that places the interests of students and teachers first:

> Any system of assessment should satisfy general criteria. For the purpose of national assessment we give priority to the following four criteria:
> - the assessment results should give direct information about pupils' achievement in relation to objectives: they should be criterion-referenced;
> - the results should provide a basis for decisions about pupils' further learning needs: they should be formative;
> - the grades should be capable of comparison across classes and schools . . . so the assessments should be calibrated or moderated;
> - the ways in which criteria are set up and used should relate to expected routes of educational development, giving some continuity to a pupil's assessment at different ages: the assessments should relate to progression.[37]

The task is to define *reliable assessment* in a different way, committing or reallocating the time and money needed to obtain more authentic and equitable tests within schools. As the British proposals imply, the professionalization of teaching begins with the freedom and responsibility to set and uphold clear, appropriate standards – a feat that is impossible when tests are seen as onerous add-ons for "accountability" and are designed externally (and in secret) or administered internally in the last few days of a semester or year.

The redesign of testing is thus linked to the restructuring of schools. The restructuring must be built around intellectual standards, however, not just around issues involving governance, as has too often been the case so far. Authentic restructuring depends on continually asking a series of questions: What new methods, materials, and schedules are required to test and teach habits of mind? What structures, incentives, and policies will insure that a school's standards will be known, reflected in teaching and test design, coherent schoolwide, and high enough but still reachable by most students? Who will monitor for teachers' failure to comply? And what response to such failure is appropriate? How schools frame diploma requirements, how the schedule supports a school's aims, how job descriptions are written, how hiring is carried out, how syllabi and exams are designed, how the grading system reinforces standards, and how teachers police themselves are all inseparable from the reform of assessment.

Authentic tests must come to be seen as so essential that they justify disrupting the habits and spending practices of conventional schoolkeeping. Otherwise standards will simply be idealized, not made tangible. Nor is it "soft-hearted" to worry primarily about the interests of students and teachers: reform has little to do with pandering and everything to do with the requirements for effective learning and self-betterment. There are, of course, legitimate reasons for taking the intellectual pulse of students, schools, or school systems through standardized tests, particularly when the results are used as an "anchor" for school-based assessment (as the British propose). But testing through matrix sampling and other less intrusive methods can and should be more often used.

Only such a humane and intellectually valid approach to evaluation can help us insure progress toward national intellectual fitness. As long as we hold simplistic monitoring tests to be adequate models of and incentives for reaching our intellectual standards, student performance, teaching, and our thinking and discussion

about assessment will remain flaccid and uninspired.

1. For an explanation of the state reports of above-average test scores, see Daniel Koretz, "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?," *American Educator*, Summer 1988, pp. 8-15, 46-52; and Edward Fiske, "Questioning an American Rite of Passage: How Valuable Is the SAT?," *New York Times*, 1 January 1989.

2. Norman Frederiksen, "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist*, vol. 39, 1984, p. 200.

3. Grant Wiggins, "Rational Numbers: Scoring and Grading That Helps Rather Than Hurts Learning," *American Educator*, Winter 1988, pp. 20, 25, 45, 48.

4. Raymond Callahan, *Education and the Cult of Efficiency* (Chicago: University of Chicago Press, 1962), pp. 80-84.

5. David Tyack, *The One Best System: A History of American Urban Education* (Cambridge, Mass.: Harvard University Press, 1974), pp. 140-46.

6. Callahan, pp. 100-101.

7. Richard J. Stiggins, "Revitalizing Classroom Assessment: The Highest Instructional Priority," *Phi Delta Kappan*, January 1988, pp. 363-68.

8. Peter Elbow points out that in all performance-based education, the teacher goes from being the student's adversary to being the student's ally. See Peter Elbow, *Embracing Contraries: Explorations in Teaching and Learning* (New York: Oxford University Press, 1986).

9. For more on content as knowledge in use and on the design of curricula and tests around "essential questions," see Grant Wiggins, "Creating a Thought-Provoking Curriculum," *American Educator*, Winter 1987, pp. 10-17.

10. Gilbert Ryle, *The Concept of Mind* (London: Hutchinson Press, 1949).

11. M. McCloskey, A. Carramaza, and B. Green, "Naive Beliefs in 'Sophisticated' Subjects: Misconceptions About Trajectories of Objects," *Cognition*, vol. 9, 1981, pp. 117-23.

12. See also Walter Haney, "Making Testing More Educational," *Educational Leadership*, October 1985, pp. 4-13.

13. Robert Glaser, "Cognitive and Environmental Perspectives on Assessing Achievement," in Eileen Freeman, ed., *Assessment in the Service of Learning: Proceedings of the 1987 ETS Invitational Conference* (Princeton, N.J.: Educational Testing Service, 1988), pp. 40-42; and idem, "The Integration of Instruction and Testing," in Eileen Freeman, ed., *The Redesign of Testing for the 21st Century: Proceedings of the 1985 ETS Invitational Conference* (Princeton, N.J.: Educational Testing Service, 1986).

14. Frederiksen, p. 199.

15. For a complete account of the nine "Common Principles," see Theodore R. Sizer, *Horace's Compromise: The Dilemma of the American High School*, updated ed. (Boston: Houghton Mifflin, 1984), Afterword. For a summary of the idea of "exhibitions," see Grant Wiggins, "Teaching to the (Authentic) Test," *Educational Leadership*, April 1989.

16. I wish to thank Albin Moser of Hope High School in Providence, R.I., for this example. For an account of a performance-based history course, including the lessons used and pitfalls encountered, write to David Kobrin, Department of Education, Brown University, Providence, RI 02912.

17. I wish to thank Dick Esner of Brighton High School in Rochester, N.Y., for this example. Details on the ground rules, the information supplied for the simulation, the logistics, and the evaluation can be obtained by writing to Esner.

18. Manuals are available from the Office of Research and Evaluation, Connecticut Department of Education, P.O. Box 2219, Hartford, CT 06115. For further information on the ACTFL guidelines and their use, see *ACTFL Provisional Proficiency Guidelines* (Hastings-on-Hudson, N.Y.: American Council on the Teaching of Foreign Languages, 1982); and Theodore Higgs, ed., *Teaching for Proficiency, the Organizing Principle* (Lincolnwood, Ill.: National Textbook Co. and ACTFL, 1984).

19. See Paul LeMahieu and Richard Wallace, "Up Against the Wall: Psychometrics Meets Praxis," *Educational Measurement: Issues and Practice*, vol. 5, 1986, pp. 12-16; and Richard Wallace, "Redirecting a School District Based on the Measurement of Learning Through Examination," in Freeman, *The Redesign of Testing . . .*, pp. 59-68.

20. Daniel P. Resnick and Lauren B. Resnick, "Standards, Curriculum, and Performance: A Historical and Comparative Perspective," *Educational Researcher*, vol. 14, 1985, pp. 5-21.

21. Aristotle *Nicomachean Ethics* 1137b25-30.

22. *Learning by Doing: A Manual for Teaching and Assessing Higher-Order Thinking in Science and Mathematics* (Princeton, N.J.: Educational Testing Service, Report No. 17-HOS-80, 1987).

23. Similar work on a research scale is being done in the U.S. as part of what is called "diagnostic achievement assessment." See Richard Snow, "Progress in Measurement, Cognitive Science, and Technology That Can Change the Relation Between Instruction and Assessment," in Freeman, *Assessment in the Service of Learning . . .*, pp. 9-25; and J. S. Brown and R. R. Burton, "Diagnostic Models for Procedural Bugs in Basic Mathematical Skills," *Cognitive Science*, vol. 2, 1978, pp. 155-92.

24. *Mathematical Development, Secondary Survey Report No. 1* (London: Assessment of Performance Unit, Department of Education and Science, 1980), pp. 98-108.

25. Task Group on Assessment and Testing, *(TGAT) Report* (London: Department of Education and Science, 1988), Paragraphs 73-75.

26. Personal communication from Joan Baron, director of the Connecticut Assessment of Educational Progress.

27. Benjamin Bloom, George Madaus, and J. Thomas Hastings, *Evaluation to Improve Learning* (New York: McGraw-Hill, 1981), pp. 52-53.

28. Jeannie Oakes, *Keeping Track: How Schools Structure Inequality* (New Haven, Conn.: Yale University Press, 1985), pp. 10-13.

29. Ibid.

30. On the engaging quality of "exhibitions" of mastery, see Sizer, pp. 62-68.

31. For various group testing and grading strategies, see Robert Slavin, *Using Student Team Learning*, 3rd ed. (Baltimore: Johns Hopkins Team Learning Project Press, 1986).

32. Credit for some of these criteria are due to Arthur Powell, Theodore Sizer, Fred Newmann, and Doug Archbald and to the writings of Peter Elbow and Robert Glaser.

33. Glaser, "Cognitive and Environmental . . . ," pp. 38-40.

34. See the work of the ARTS Propel project, headed by Howard Gardner, in which the portfolio idea is described as it is used in pilot schools in Pittsburgh. ARTS Propel is described in "Learning from the Arts," *Harvard Education Letter*, September/October 1988, p. 3. Many members of the Coalition of Essential Schools use portfolios to assess students' readiness to graduate.

35. Alverno College has prepared material on the hows and whys of self-assessment. See Faculty of Alverno College, *Assessment at Alverno*, rev. ed. (Milwaukee: Alverno College, 1985).

36. For a lively discussion of the research results on the special ETS testing conditions for dyslectics, who are given unlimited time, see "Testing, Equality, and Handicapped People," *ETS Focus*, no. 21, 1988.

37. Task Group on Assessment and Testing, *(TGAT) Report* (London: Department of Education and Science, 1988), Paragraph 5.

BEST COPY AVAILABLE

Reprinted with permission from EDUCATIONAL LEADERSHIP, February 1993.

How World-Class Standards Will Change Us

Arthur L. Costa

As we abandon traditional views of education, the skills of thinking and problem solving will replace discrete subject areas as the core of the curriculum and will lead to changes in instruction and assessment.

Enterprising learning communities looking toward education in the next century know that new educational goals are imperative for our youth's survival, the continuance of our democratic institutions, and even for our planetary existence. Learning how to learn throughout a lifetime; knowing how to behave when answers to complex problems are ambiguous, dichotomous, and paradoxical; and generating, organizing, and applying an abundance of technological information are just a sample of the types of goals we need to establish.

From such goals, we will establish world-class standards. But to do this, we must be prepared for a paradigm shift. We will have to replace some of our obsolete, traditional views of education with more modern, relevant, and consistent ones. We will let go of our obsession with content acquisition and knowledge retention as merely ends in themselves. We will dismiss uniformity and begin to value diversity. We will extinguish external evaluation of students and teachers in favor of self-evaluation. We will replace extrinsic rewards with learning activities that are intrinsically motivating. We will deflate competitiveness to expand interdependence. We will redefine *smart* to mean knowing how to draw forth from a repertoire of strategies, knowledge, perceptions, and actions according to contextual demands.

Since all aspects of a system are interlocking, all parts must change in accordance with the new paradigm. No one part can operate efficiently unless the other parts of the system work harmoniously (Kuhn 1970). Imposing higher standards, therefore, is not just "kid stuff." Higher standards must be set for all components of the educational enterprise. We will need to evaluate the contribution of all parts of the system to higher standards in curriculum, instruction, and assessment.

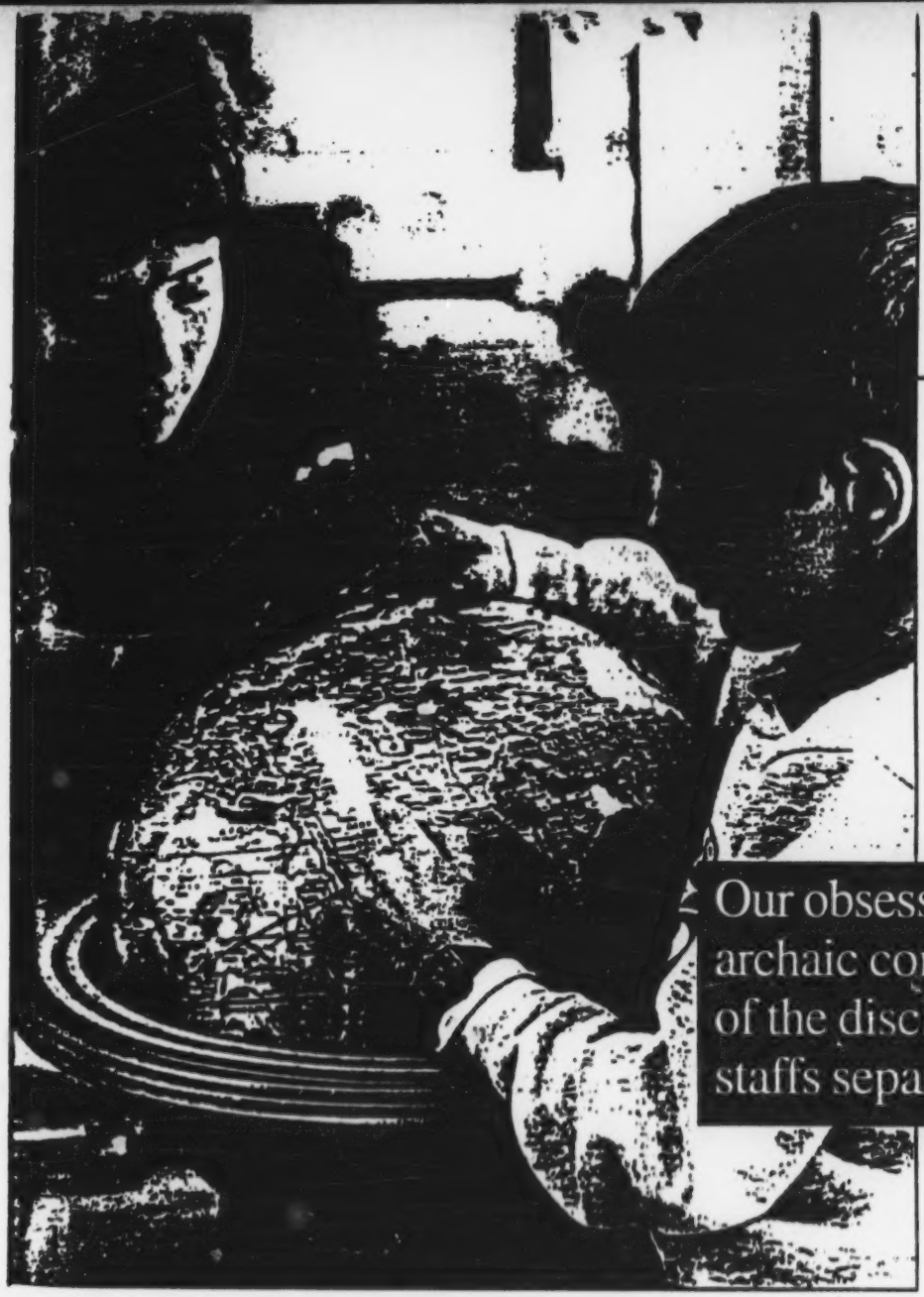
Higher curriculum standards. Our obsession with the archaic compartmentalization of the disciplines keeps school staffs separated. As intellectual development, thinking, problem solving, and cooperating become the core of the curriculum, traditional content and subject-matter boundaries will become increasingly obscure and selectively abandoned. Process will become the content of instruction.

Our view of instruction will shift from learning *of* the content to learning *from* the content. History, physical sciences, mathematics, or the arts in the curriculum will no longer be ends in themselves; instead, problem-based instruction will serve as the vehicle for learners to integrate, reflect on, and transfer the unique knowledge, structure, and modes of inquiry of several disciplines. Teachers will select relevant, problem-centered, integrative themes because of their contributions to the thinking/learning process, and we will focus standards on applying concepts from a variety of fields to produce new knowledge, transfer strategies to novel situations, and tackle complex problems.

Higher instructional standards. Having world-class standards will make us begin to see the profession of teaching as intellectually complex, collaborative, and reflective. Teachers who have achieved the highest stages of intellectual functioning are more committed to and empathic with individual students, and they provide greater stimulation for students to function at higher levels of cognitive complexity as well (Glickman 1985).

In traditional school settings, time limits, isolation, and minimal peer interaction prevent teachers of different departments, grade levels, and disciplines from meeting together. Thus, teachers' intellectual growth is diminished. Teaching toward world-class standards will require mutual interaction from teachers who work with students from various age levels and possess rich funds of knowledge in various fields. New roles for members of instructional teams will be:

- *knowledge managers* who judiciously select topics and problems for their contribution to achieving these standards;
- *team planners* who continually clarify the school community's desired outcomes and create instructional strategies to achieve their goals;
- *collegial coaches* who teach, observe, and give nonjudgmental feedback to one another;
- *collaborative researchers* who experiment with and evaluate curriculum and instructional effectiveness and modify them accordingly.



Our obsession with the archaic compartmentalization of the disciplines keeps school staffs separated.

Higher assessment standards. We cannot employ traditional product-oriented assessment techniques to evaluate the achievement of these new, process-oriented standards. Skillful teaching teams and students themselves (who are the best collectors of assessment data) will, over time, observe, record, interpret, and report evidence of growth toward and achievement of these standards. Such assessment strategies will include:

- directly observing performance in collaborative problem-solving situations;
- collecting logs, journals, and portfolios of selected artifacts of learning excellence;
- observing performances while conducting extended cooperative projects;
- conducting interviews to discover students' self-perceptions as problem solvers;
- maintaining checklists recording indicators of growth toward desirable habits of mind;
- assessing displays, exhibitions, and performances according to both internal and external criteria; and
- employing media and advanced technology to assist in collecting and recording information.

The entire community must become committed to higher standards of educational excellence. This means higher standards of social services as well: health care, employment, child care, recreation, continuing education, housing, and welfare. The problems children bring to school that detract from their achievement of higher standards are often societal, not learning, problems. Schools, being a reflection of society, will achieve higher, world-class standards only when American society imposes higher, world-class standards on itself. Achieving higher standards requires the devotion of the greatest share of our resources to the development of each person's fullest potential. ■

References

Glickman, C. (1985). *Supervision of Instruction: A Developmental Approach.* Newton, Mass.: Allen and Bacon.

Kuhn, T. (1970). *The Structure of Scientific Revolutions, International Encyclopedia of Unified Science.* Chicago: University of Chicago Press.

Arthur L. Costa is Professor Emeritus, California State University, Sacramento. He can be reached at P. O. Box 705, Kalaheo, HI 96741.

Reprinted with permission from TEACHER Magazine,
Volume IV, Number 6, March 1993.

INNOVATIONS

Smart Tests

A new approach to assessment helps teachers understand the ways young children think and grow.

When Mary Ann Cockman asks her kindergarten pupils to draw self-portraits at the start of the school year, they often sketch a large head, two eyes, and a stick for a body. When she asks them to do the same thing later in the year, their renderings become more detailed, acquiring ears, mouth, nose, eyebrows, eyelashes, fingers, shoes, rectangular legs—even designs on their clothing.

Cockman, who teaches at Thomson Elementary School in Davidson, Mich., says the student portfolios she now collects to document such progress—as well as the detailed checklists and written evaluations she prepares several times a year on each child—have helped her become more attuned to the ways young children think and grow. "You learn to look at children differently," she says.

Thomson Elementary is one of 37 schools nationwide piloting a system of assessment designed to provide an alternative to standardized tests and traditional report cards in the early grades. The approach, known as the "Work Sampling System," was developed by Samuel Meisels, a professor of education and an associate dean for research at the University of Michigan.

Meisels is one among many experts who argue that traditional standardized tests are unreliable for young children and have been used inappropriately to delay school entry, retain large numbers of students, and justify rote-skill instruction. Meisels' goal is to replace group-administered achievement testing with teacher-focused performance assessment. Such a change, he says, would improve both teaching practices and children's experiences in the early grades.

Teachers using the Work Sampling System complete a detailed checklist three times a year tracking children's performance in seven domains: personal and social development, language and literacy, mathematical thinking, scientific thinking, social and cultural awareness, art and music, and physical development. They also collect samples of children's work in each domain, including tasks asked of all children and individual items that vary by child.

Then, using the checklists, portfolios, and other observations, the teachers complete a "summary report," or student profile, evaluating a child's performance, strengths, and difficulties in each domain. These reports, which are also prepared three times a year, are designed to serve as a basis for communicating with other teachers, parents, and administrators and for planning student instruction.

Initial studies to determine whether outside experts would rate children the same way their classroom teachers do based on the portfolios and checklists have shown "extremely high" reliability, according to Meisels. And a study designed to gauge the validity of the assessment compared with an individually administered norm-referenced standardized test, he says, also produced promising results.

Still, Meisels cautions that the Work Sampling System should not be used as a "gatekeeping device" for school entry or promotion. And while data derived from the system could be used to compare the performance of classrooms or large groups, he says, its chief purpose is to assess individual children's progress over time, gauge achievement against curriculum goals, and tailor instruction to children's needs.

Lorrie Shepard, a professor of education at the University of Colorado and author of studies on the adverse effects of retention in the early grades, says the biggest danger in introducing alternative early childhood assessments occurs "when people take a measure designed for one purpose and retrofit it for another." Much more experimentation and fieldwork in the higher grades are needed, she maintains, to reliably use such data to rate the quality of teachers and schools. "The kinds of preliminary data we've seen are sufficient to justify classroom uses of these assessments," she says, "but would not be sufficient for external accountability purposes."

Such assessments are perhaps most useful in helping teachers learn more about child development. JoAnne Lowe, a kindergarten teacher at Copeland Elementary, a pilot school in Dexter, Mich., says the Work Sampling System has helped her become a better observer of the whole gamut of children's needs—including their physical growth and personal and social development—and has "made me aware of areas I wasn't strong at."

"One of the best things about it is that children do not fail," says Janice Brown, principal of Kettering Elementary in Willow R... Mich., another participati... school. "What you are measuri... a child against is him- or hers... not a false presumption like... grade or test. If no growth is e... dent, then we can look at it a... look at how we are going... change our system."

One drawback, teachers say... that the system is very 'on... consuming. But, they add, p... ents like the detailed feedb... this kind of assessment gi... them, and it buoys children's s... esteem. "Kids, by the end of t... year, are able to look at th... portfolio and see their o... growth," says Penny Butler, a k... dergarten teacher at Garfield E... mentary School in Flint, Mich.

While they applaud the level... training and technical assistar...

> 'One of the best things about it is that children do not fail,' says on... principal. 'What you are measurin... a child against is him- or herself, not a false presumption like... a grade or test.'

provided by Meisels and his co... leagues, some educators still pr... dict it will take a few years f... them to become comfortabl... using the system. And Harvar... University psychologist Howar... Gardner notes that it also wi... take time to "develop canons ... looking and listening"—like thos... established to rate performanc... in other fields—that pol... cymakers would consider accep... able for accountability purposes.

Says Principal Janice Brow... "One of the hard questions educ... tors have to answer is: Are w... interested in knowing how chi... dren compare with one anothe... or are we interested in knowin... how much children are learning... If we have decided we are mor... interested in knowing how muc... children are learning, we must d... something like this." ■

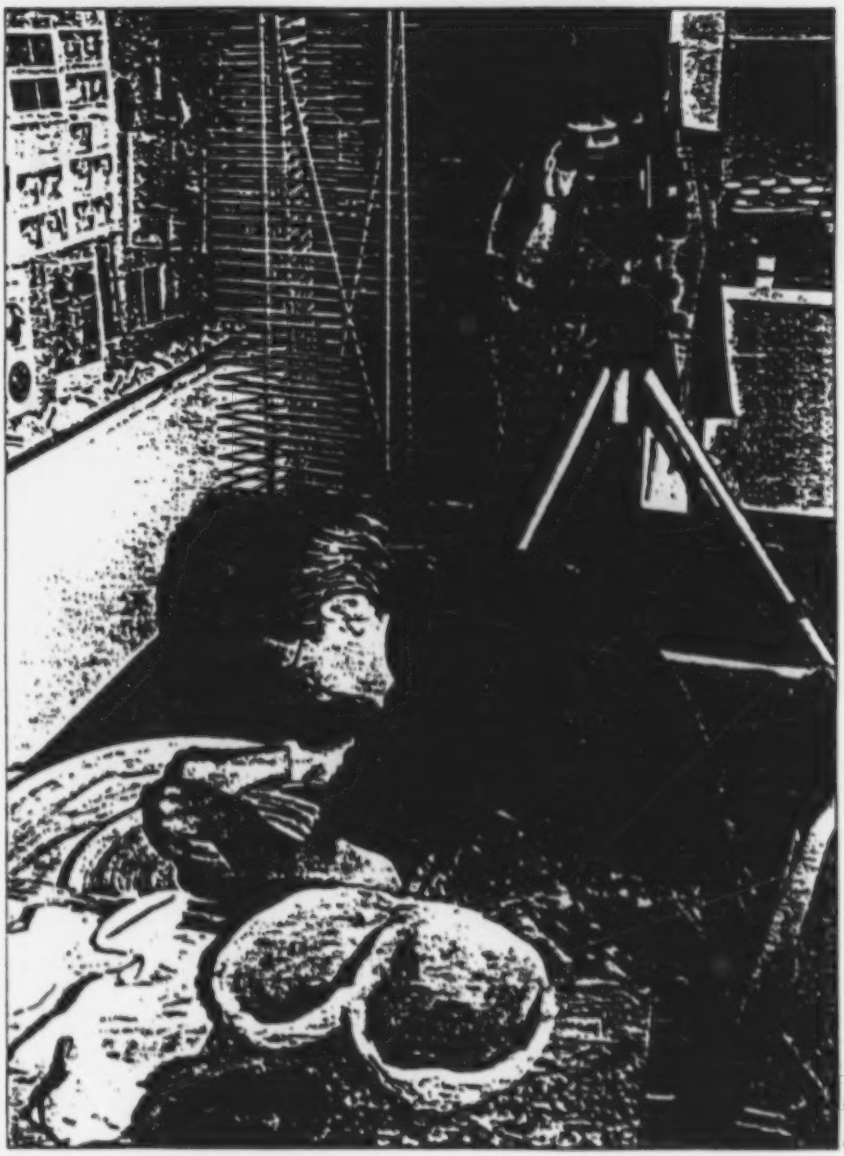
—Deborah L. Cohe...

BEST COPY AVAILABLE

Reprinted with permission from Jo Campbell, Principal, Gillette, Wyoming.

Laser Disk Portfolios: Total Child Assessment

Multimedia technology has transformed the assessment process for students and teachers at a rural elementary school in Wyoming.



JO CAMPBELL

As this story begins, a 4th grade student is sitting at a multimedia system consisting of a computer, CD-ROM drive, optical drive, scanner, and laser printer. The student has requested that a story he wrote become part of his permanent assessment record. He is scanning his work onto his laser disk. As he works, another student asks to take the video camera to the art room. She wants to record a video of herself throwing a pot on the potter's wheel as one of her assessments for the year. When she finishes, she returns to the multimedia system to transfer the video to her laser disk.

Earlier today, a student has used the video camera to record the play in which he had the leading role. Another student has documented on disk that she has the ability to climb the rope in physical education class.

All of this is happening at Conestoga Elementary School in rural Wyoming, where the idea of creating portfolios of students' work as a means of assessment has been combined with laser disk technology. The laser disk portfolio assessment system is based on the work of IBM consultants and researchers from Project Zero at Harvard. A grant from the Wyoming State Department of Education Super School Program helped our school purchase the necessary IBM hardware. While the technical implementation has had its frustrating moments, the IBM consultants and school staff have persevered.

Using Laser Disk Assessments

Students and teachers at Conestoga have begun to use the laser disk portfolio assessment system this year. Large amounts of information can be added to or retrieved from the system as many times as necessary. Yet the laser disks are so small they can slide into any student's permanent file, eliminating the need to find additional filing space.

Teachers can use the system to research their classes before the first day of school. The system helps them do something as simple as putting students' names and faces together, or something as complicated as deciding what teaching and learning activities would best increase outcomes for the students.

To begin planning assessment, grade-level teams determine criteria to evaluate growth for each child. Teachers do a preassessment the first month of school by videotaping their students. These preassessment videos might record a student playing a game with another student, reading a story of the student's choice, or answering a few questions as the video camera records the student's ability to talk with an adult.

Teachers plan to use the system for formal assessments at least twice a year with each student. Has the student gained reading fluency? Has handwriting improved? How else has the child grown? Questions such as these can be answered by reviewing the disk.

During the school year, teachers can review the disks of students who are struggling with schoolwork to determine what the students' interests are and in what areas they have excelled previously. The disks can be used to motivate the child who says, "I can't do it," by showing how much he or she has done over the years. Self-esteem builds as the accomplishments of each child are recorded and as growth is measured against individual standards, rather than group standards.

At the end of the school year, an annual ritual of passage will take place as each 6th grader and his or her parents, along with the teacher and principal, review the student's disk from kindergarten to 6th grade. We are expecting tears and laughter and possibly embarrassment from students viewing their antics of years gone by, but we are also predicting that every family will want a copy of the disk for a personal history. In any event, the school will keep a disk for its records.

Implementing the System

A school planning team handled the implementation of Conestoga's laser disk portfolio assessment system. The counselor made home visits to talk to parents about the system and ascertain any concerns or ideas about assessing in this manner. Implementing the complete system with just one class the first year allowed staff participants to acquaint themselves with the system and become comfortable with the equipment before launching into assessments of the total school population.

The kindergarten class will be the first class to be followed through the grades. This year we videotaped the kindergartners answering predetermined questions and transferred that image to their laser disks. Teachers also scanned the children's drawings onto the disks.

Getting a Complete Picture

Portfolio assessment is not a new concept in education, but a system that allows permanent storage of optical data, written and drawn images, and verbal ability *is* new. The staff and parents at Conestoga are excited about this new assessment system, especially as it relates to increasing student self-esteem. Recognizing that child development involves more than cognitive growth is the underlying belief behind the system. The students, parents, and staff will be looking at the laser disk portfolio assessments to determine growth in verbal ability, physical accomplishment, artistic achievement, and self-assurance. □

Copyright © 1992 by Jo Campbell.

Jo Campbell is Principal of Conestoga Elementary School and a doctoral candidate at Teachers College, Columbia University, New York City. She can be contacted at Conestoga Elementary School, 4901 Sleepy Hollow Blvd., Gillette, WY 82716.

Reprinted with permission from EDUCATIONAL LEADERSHIP, May 1992.

Portfolios Invite Reflection — from Students *and* Staff

Learning Experience Forms and Portfolio Evenings are just two of the ways Crow Island School found to make the assessment process more meaningful.

ELIZABETH A. HEBERT

Four years ago, Crow Island Elementary School began a project that has reaped benefits far beyond what any of us could have imagined. The focus of the project was assessment of children's learning, and the tangible product is a new reporting form augmented by student portfolios.

More important, however, has been the process of developing our thinking and teaching around new ways of looking at children's learning. In fact, this process became more valuable to us as a faculty than the assessment product, helpful as it has been.

Our Commitment to Alternative Assessment

The project grew out of our dissatisfaction and frustration with mandated standardized modes of assessment. Standardized tests do not reflect how we teach, the effects of our teaching on children, or how we adapt instruction to individual learners. Wolf and colleagues write that "the design and implementation of alternative modes of assessment will entail nothing less than a wholesale transition from what we call a *testing* culture to an *assessment* culture." They continue:

> The observable differences in the form, the data, and the conduct of standardized testing and its alternatives are in no way superficial matters or mere surface features. They derive from radical differences in underlying conceptions of mind and of the evaluation process itself. Until we understand these differences and their network of consequences, we cannot develop new tools that will allow us to ensure that a wide range of students use their minds well (1991, p. 33).

Obviously, we had our work cut out for us. What did we do to reaffirm our commitment to a concept of learning incompatible with standardized testing? First, we did a good deal of reading; engaged in lengthy discussions about values, community building, and conferencing; and consulted with experts. We also became more deliberate about making time to visit one another's classrooms and to share and refine our observations of children. Next we began defining the questions to which we were seeking answers. Our first questions were global:

- How do we define learning?
- Where does learning take place?
- How do we recognize learning?
- How do we report instances of learning?

As we answered these larger questions, our concerns became more specific. How can we communicate about children's learning experiences with parents in ways that:

- authentically describe the child,
- speak to issues of accountability and maintain the integrity of our beliefs about children and how they learn,
- reflect the different ways that teachers organize instruction,
- provide concrete information compatible with parents' expectations?

A Compatible Theory

Some background information about our school provides a context for our project and how we went about answering these questions. Crow Island is a public JK-5 school in Winnetka, Illinois, an affluent suburb of Chicago's north shore. The Winnetka Public Schools include three elementary schools and one middle school for grades 6-8. Our lower schools have enrollments of 360-390. Although a public school system, we have a strong tradition in the progressive philosophy of education that is distinguished by:

- a commitment to a developmental orientation to instruction,
- the priority placed on consideration of the "whole child" and his or her individual mode of learning,
- the absence of letter grades until 7th grade,
- high regard for teachers as professionals.

In acknowledging the uniqueness of a child's mode of learning, the district has placed a high priority on conferencing with parents. For many years, pupil progress has been reported to parents in a conference format three times per year. Teachers had prepared narrative descriptions of children

using the following organizers: language arts, math, social studies, science, growth of the child as a learner, and growth of the child as a group member.

One expert who influenced our thinking about alternative assessment was Howard Gardner, whose "Theory of Multiple Intelligences" (musical, linguistic, logical-mathematical, spatial, bodily-kinesthetic, interpersonal, and intrapersonal) challenges the more traditional concepts of intelligence. The main thrust of Gardner's theory as applied to schools is that children may demonstrate the different kinds of intelligences in ways not necessarily associated with traditional school subjects and certainly not associated with traditional modes of assessment. Gardner's theory resonated with the themes of progressive education to which we at Crow Island are devoted.

A Visual Format

Gardner's theory provided a good scaffold for our thinking. The next step was to put our thoughts into a visual format. Our first rough attempt began to capture the idea of multiple dimensions of a child's learning. This primitive model consisted of a stick figure surrounded by floating boxes. As you may expect, there was much discussion about the number, size, and positioning of the boxes, but we finally agreed on a format. We call it our Learning Experiences Form (see fig. 1 for a composite example of the form, shortened for space).

Our next concern was to identify our organizers on the Learning Experiences Form. Being committed to the multiple intelligences perspective, we readily included music, art, and physical education. We wanted to recognize these teachers' long-term relationships with students, the value of their programs, and their insights about children's learning. But what about the other Learning Experience organizers? The dialogue went something like this:

> Q: How should I specify my organizers?
>
> A: That depends on how you organize instruction.
>
> Q: But what if mine are different from someone else's?
>
> A: That's OK. You organize instruction differently. We already know that about one another. Now we're just writing about it.
>
> Q: But we organize instruction differently for different students.
>
> A. Your Learning Experiences Form will then reflect the flexibility of your teaching.

This was a crucial stage in our thinking because discussing the form brought to the surface what I term the "bilingualism" of teachers. *Inside language* — what we do in our classrooms — reflects our beliefs and values, years of teaching experience, observations of children and of other good teachers, and confidence in knowing what we know. *Outside language* — what we *say* we do in our classrooms — is influenced by community values, comfort level within the school environment, political pressures, district and administrative policies, test scores, and curriculum.

The nature of our project necessitated our speaking "inside language," a more difficult discourse because it requires feelings of safety and security. Gradually, though, we were able to experience the sharing of values that leads to the creation of a secure, thoughtful environment for children, teachers, and parents.

A Close Look at Ourselves

In order to change how we evaluated

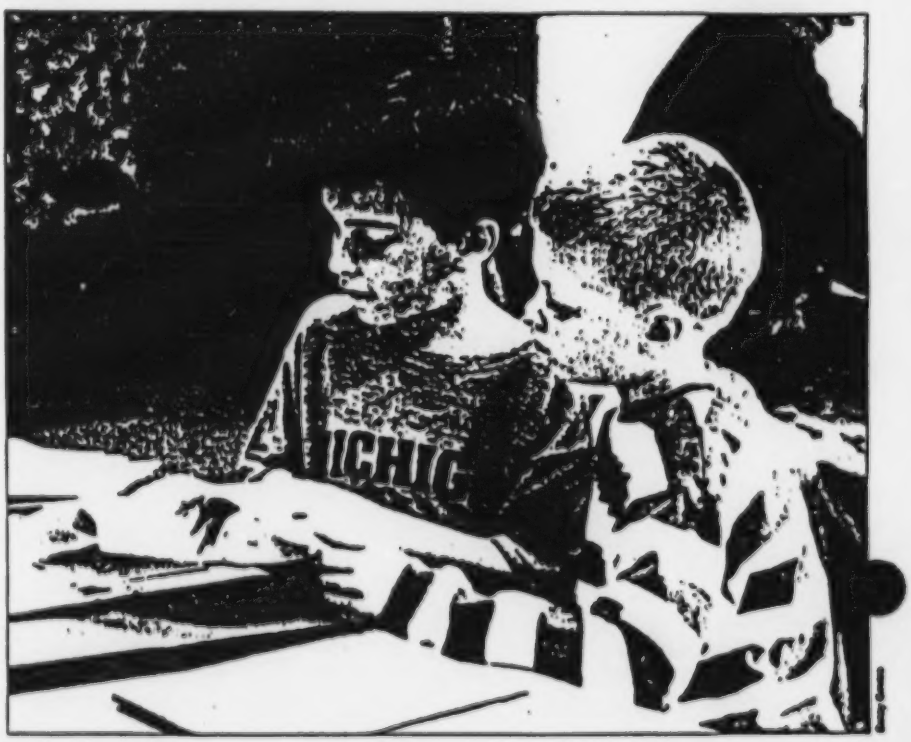
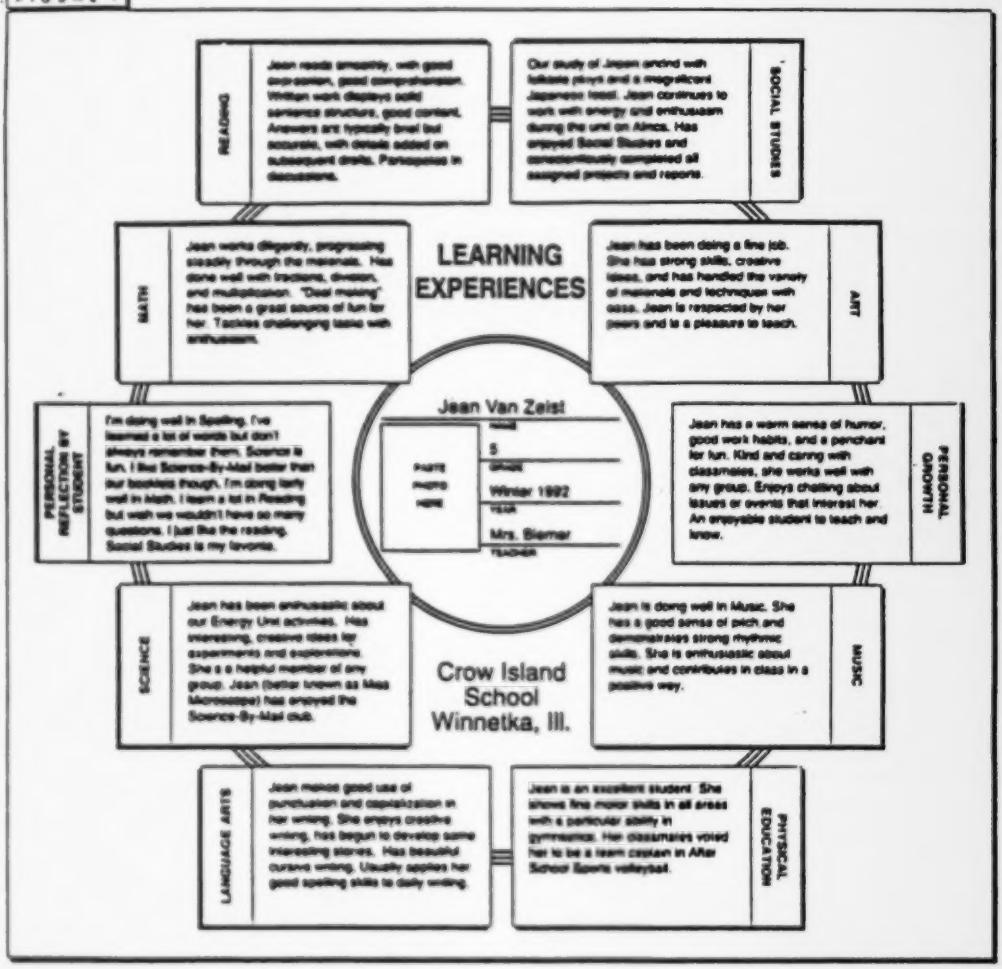


FIGURE 1


READING
Jean reads smoothly, with good expression, good comprehension. Written work displays solid sentence structure, good content. Answers are typically brief but accurate, with details added on subsequent drafts. Participates in discussions.
SOCIAL STUDIES
Our study of Japan ended with folktale plays and a magnificent Japanese feast. Jean continues to work with energy and enthusiasm during the unit on Africa. Has enjoyed Social Studies and conscientiously completed all assigned projects and reports.
MATH
Jean works diligently, progressing steadily through the materials. Has done well with fractions, division, and multiplication. "Deal making" has been a great source of fun for her. Tackles challenging tasks with enthusiasm.
LEARNING EXPERIENCES
ART
Jean has been doing a fine job. She has strong skills, creative ideas, and has handled the variety of materials and techniques with ease. Jean is respected by her peers and is a pleasure to teach.
PERSONAL REFLECTION BY STUDENT
I'm doing well in Spelling. I've learned a lot of words but don't always remember them. Science is fun. I like Science-By-Mail better than our booklets though. I'm doing fairly well in Math. I learn a lot in Reading but wish we wouldn't have so many questions. I just like the reading. Social Studies is my favorite.
Jean Van Zeist
NAME
5
GRADE
Winter 1992
YEAR
Mrs. Biemer
TEACHER
PHOTO
PERSONAL GROWTH
Jean has a warm sense of humor, good work habits, and a penchant for fun. Kind and caring with classmates, she works well with any group. Enjoys chatting about issues or events that interest her. An enjoyable student to teach and know.
SCIENCE
Jean has been enthusiastic about our Energy Unit activities. Has interesting, creative ideas for experiments and explorations. She is a helpful member of any group. Jean (better known as Miss Microscope) has enjoyed the Science-By-Mail club.
Crow Island School Winnetka, Ill.
MUSIC
Jean is doing well in Music. She has a good sense of pitch and demonstrates strong rhythmic skills. She is enthusiastic about music and contributes in class in a positive way.
LANGUAGE ARTS
Jean makes good use of punctuation and capitalization in her writing. She enjoys creative writing, has begun to develop some interesting stories. Has beautiful cursive writing. Usually applies her good spelling skills to daily writing.
PHYSICAL EDUCATION
Jean is an excellent student. She shows fine motor skills in all areas with a particular ability in gymnastics. Her classmates voted her to be a team captain in After School Sports volleyball.

children's learning, we realized we needed to take a close look at ourselves. We soon found ourselves undergoing an intensive assessment of our teaching, our beliefs about children, and our views of the school and its relationship to our community.

At this point, the project quite naturally proceeded from an emphasis on student assessment to a more powerful staff development focus. In order for this to occur in any school, administrators must commit to providing the kind of school environment where such a climate can flourish. Administrators also have to acknowledge that all teachers do not arrive at the same point in their growth together. As we emphasize with the children, teachers must construct their own knowledge of children, how they learn, and how to evaluate that learning. We have to be patient and sufficiently open to allow for different stages of understanding, yet focused enough to provide clarity and vision to the effort.

Improvements to the Process

We began using the Learning Experiences Form in a variety of ways. Some teachers were more conservative, using traditional school subject labels on their forms. Others coined new organizers that reflected their teaching styles. As they struggled with the new format, teachers became more thoughtful; and parents, sensing the positive energy and concern of teachers, responded enthusiastically. After the first conference using the new form, the response from both parents and teachers was overwhelmingly positive.

Over four years, we've refined the form to meet the suggestions of teachers at kindergarten, primary, and intermediate levels. In response to our concern about how to separate out curriculum specifics and descriptions of a child's learning, one of our teachers designed a Curriculum Overview, to be printed on the back of the form, that consists of mini-statements of curriculum objectives for that portion of the year. This addition freed up the front of the form for more focused descriptions of children's learning.

Noting the absence of the child's input to the form, we designated a space for a "child's reflection" about his or her learning. The older students write their own thoughts; teachers take dictation for the 1st graders. We've also begun to include parents' thoughts about their child's learning experience in our assessment form.

Students Tell Their Stories

The next step was to have our students create portfolios. Portfolios are compatible with Crow Island's agenda for effective teaching, authentic assessment, and faculty growth. One of the best definitions in the current literature comes from Paulson and Paulson (1991): "Portfolios tell a story . . . put in anything that helps tell the story." With these authors, we also agree about the importance of the child's participation in selecting the contents of the portfolio and with a focus more on process than on content (1991, p. 1).

At present, each of our students has a portfolio that represents work across all domains. Students maintain their portfolios all year and frequently have conferences with the teacher about works in progress, additions, and deletions. At the end of the year, their portfolios are combined with past years' work and stored in our Student Archives. The archives are alphabetically arranged in open shelving in our Resource Center along with historical documents, publications, and photographs of our school and students.

Portfolio Evenings

Three years ago we added a new element to our assessment project. Encouraged by the kinds of thinking that children had expressed in their Student Reflections, we realized that they were capable of much more. Getting them more involved in the process of assessment seemed to make good sense.

In preparation for "Portfolio Evenings," children review their portfolio/archive as teachers guide them with questions like:

- How has your writing changed since last year (or since September)?
- What do you know about numbers now that you didn't know in September?
- Let's compare a page from a book you were reading last year and a book you are reading now and include copies of each in your portfolio (an idea from Denise Levine, Fordham University, New York).
- What is unique about your portfolio?
- What would you like Mom and Dad to understand about your portfolio? Can you organize it so it will show that?

The idea is to ask guiding questions that help children reflect on their learning. Students are encouraged to write about their learning and to include these thoughts as part of their portfolios. Developing the metacognitive process in students, even at a young age, heightens their awareness and commitment to a critical assessment of their learning.

In preparation for Portfolio Evenings, the teacher divides the class into small groups of six or seven at the primary level (and larger groups at grades 4 and 5) and assigns a night for each group of students and their parents. Primary-level Portfolio Evenings are held in February. We hold intermediate-level Portfolio Evenings in May, because older students prepare more extensive projects.

On Portfolio Evenings, which last for about an hour and a half, the children sit with their parents and present their portfolios. The teacher and I circulate, visiting each student and highlighting particular milestones each youngster may have attained. We are available for questions but try not to intrude, because this is really the children's evening, and they need to "run the show" as much as possible. Parents and teachers have been impressed with the leadership and independence that even our youngest students have demonstrated in this setting.

A Powerful Learning Experience

We are continuing to refine our assessment project. Some issues we're addressing are practical in nature, for example, storage containers for the portfolios/archives. Others are more fundamental, like how to use portfolios to link children's early, strong expressions of interest in a particular topic to more sophisticated elaborations later in their school careers. We are also contemplating how to gain the community's support for these alternative modes of assessment as part of a viable system of accountability. And, finally, as a faculty we are trying to preserve the cohesive and bold spirit that nurtured this project along its way.

The entire process has been a powerful learning experience for our faculty as well as for the children and their parents. It has expressed the fundamental values of our school district and represents our joint exploration of the complex issues of children and their learning. We are encouraged to go forward by the positive effects this project has had on the self-esteem and professionalism of the individual teachers and the inevitable strengthening of the professional atmosphere of the entire school. We have improved our ability to assess student learning. Equally important, we have become, together, a more empowered, effective faculty. □

References

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences.* New York: Basic Books.

Paulson, F. L., and P. R. Paulson. (Copyright, February 1991). "Portfolios: Stories of Knowing." Pre-publication draft.

Paulson, F. L., P. R. Paulson, and C. Meyer. (1991). "What Makes a Portfolio a Portfolio?" *Educational Leadership* 48, 5: 60-63.

Wolf, D., J. Bixby, J. Glen, and H. Gardner. (1991). "To Use Their Minds Well: Investigating New Forms of Student Assessment." In *Review of Research in Education* 17, edited by G. Grant, p. 33. Washington, D.C.: American Educational Research Association.

Author's note: For further reading on student archives, consult the writings of Pat Carini and teachers from the Prospect School in Bennington, Vermont, The Prospect Archive and Center for Education and Research, Bennington, VT 05257.

Elizabeth A. Hebert is Principal, Crow Island School, 1112 Willow Rd., Winnetka, IL 60093.

Reprinted with permission from THE SCHOOL ADMINISTRATOR, December 1992.

ASSESSMENT REVOLUTION

PORTFOLIO ASSESSMENT
IN THE HANDS OF TEACHERS

Skills and Training Needed To Meet a High-Stakes Challenge

BY CLARE FORSETH
Teacher, Marian Cross Public School, Norwich, Vermont

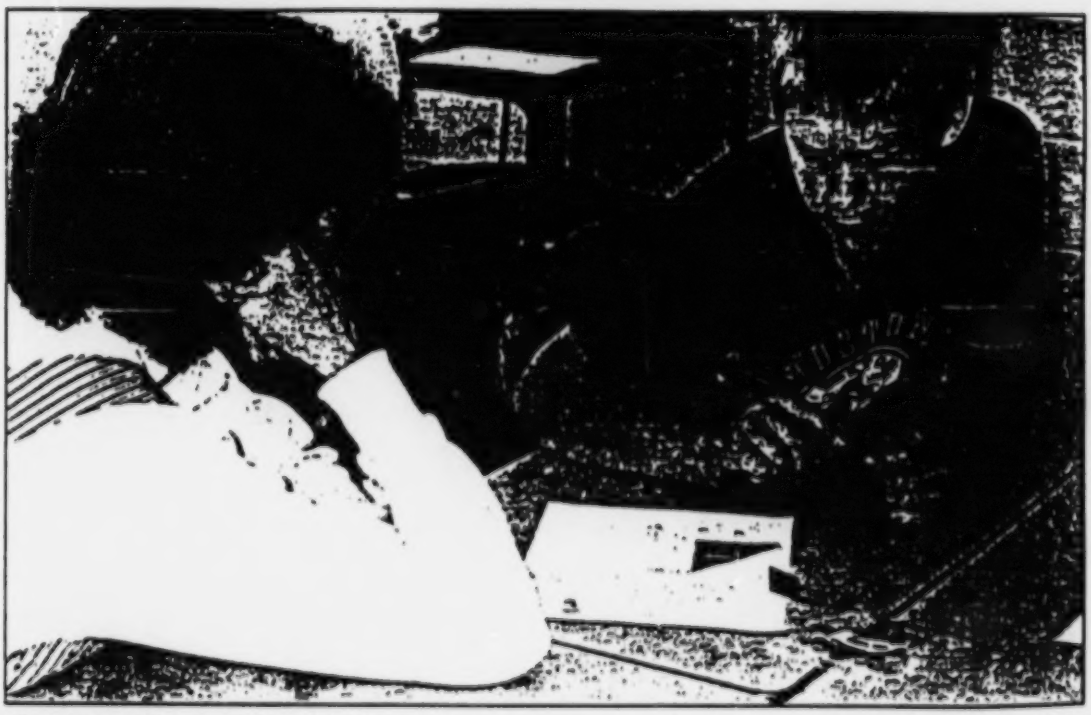
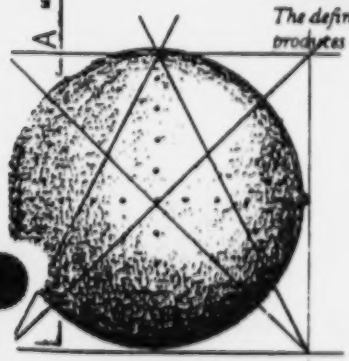


Photo by Sarah Akers

The definition of "assessment" is broader than tests. In time, educators may realize that "assessment" includes every activity that produces information about students.



Those who shape educational policy in Vermont are engaged in a formidable task: a statewide innovative assessment of mathematics and writing ability of fourth and eighth graders.

After Vermont's legislature funded a program in 1990 that allowed the state's Department of Education to investigate various assessment systems, a committee of teachers and consultants looked at what the current methods of assessment were measuring. The committee decided the long-used standardized test was not assessing many of the educational outcomes Vermonters value most.

These policymakers then turned to

Photos by Sarah Meyers

Vermont's commitment to portfolio assessment requires teachers to become more than familiar with a new assessment tool. It also demands new instructional techniques.

"To understand... teachers need to be problem solvers themselves."

Sixth-grade students from Marion Cross Public School, Norwich, Vt., are solving a mathematical riddle in which they try to find a number that could be made into a square and a cube.

a portfolio assessment system that looks at student work and assesses it against established criteria. Unlike standardized test results, which reveal what a student does and does not know on one specific occasion, the portfolio showcases the student's best work over time.

This alternative means of assessment represents a high-stakes challenge for the state's teachers, asking much more of them than mere familiarity with a new assessment tool. Successful implementation of this form of assessment calls on teachers to make significant changes in their instructional techniques.

Process Approach

To effect such changes, teachers need a variety of support systems to develop their understanding of portfolio assessment and the new skills that support it. Vermont has established support systems to develop portfolio assessment skills in mathematics and writing. The change from other assessment methods is most clearly demarcated in the mathematics portfolio assessment.

Many teachers are familiar with the process approach to writing instruction, which has been in place for more than a decade. In this approach, student writers develop their ideas through many drafts, with less emphasis on the end product than on the learning and thinking process that creates an increasingly precise refinement of ideas.

During much of this same period, though, mathematics instruction continued to be driven by standardized tests that emphasized computation and one-step word problems. The most widely used textbooks also reflected these emphases. Unfortu-

BEST COPY AVAILABLE

nately, these systems continued the stress on successful computation and correct answers that has long been the focus of the mathematics classroom, and this focus retarded the development of process-oriented approaches to teaching mathematics.

Vermont's investigation of various assessment systems convinced planners of the importance of a more thoughtful analysis of which mathematical skills and abilities are truly valuable and lasting, and how these abilities might best be fostered through instruction and assessment. They recognized the need to begin to focus on problem solving and on the students' ability to communicate their reasoning.

So significant a change in approach to assessment, combined with necessary changes in instructional approach, could not be implemented overnight, even in a small state like Vermont. Teachers first had to learn just what was to be evaluated in portfolio assessment, and how.

Evaluation Criteria

In the portfolio assessment of mathematics, student work is used to assess problem solving and communication skills using these seven criteria:

- understanding the task,
- applying the strategy,
- making decisions,
- verifying the solution,
- making connections,
- using rich mathematics language, and
- using effective mathematical representation.

The first task, then, was to familiarize teachers with these criteria.

To this end, a committee of teachers developed institutes to introduce participants to the problem solving and communication criteria used to assess student work. Teachers responded eagerly to the chance to make meaningful changes in their approach to teaching mathematics.

One participant in a portfolio institute said: "I feel both excited and challenged by the mathematics portfolio project we are undertaking as a state. It is exciting to examine alternative assessment measures and challenging to confront the reality that this assessment project is driving a change in our way of providing mathematics instruction to students. The mathematics portfolio requires some dramatic shifts in our teaching. Teachers who have traditionally used drill activities and worksheets will need to rethink their instruction in order to ensure that students become effective problem solvers, hence meeting the criteria set by the portfolio project."

To understand the process their students go through as they solve a problem, teachers need to become problem solvers themselves. During the institutes, teachers not only learn to assess student work, they also are actively engaged in problem solving. Sample problem-solving situations are presented to the teachers. One day teachers use tangram pieces to investigate fractions. Another day they use Cuisenaire rods to introduce the concept of ratio and proportion. In each case the teachers work in cooperative groups and write up and share their strategy and reasoning with other participants. This type of model lesson allows the teacher to see exemplary instructional techniques as they experience being problem solvers.

Student Ownership

Teachers who have participated in a portfolio institute come to understand problem solving involves more than simply getting the answer. It also encompasses communicating through mathematical language and a representation of how the student solved the problem, why certain decisions were made along the way, and what connections the student was able to make during the problem-solving process.

An important goal of the institutes is for teachers to convey this awareness of the problem-solving process to students in their mathematics classes. This means involving students in many problem-solving situations and assessing the work using problem-solving and communication criteria.

Through the assessment, students begin to understand how to improve their solutions or presentations. Students who take ownership of the criteria begin to be able to assess their own work. They begin to focus on aspects of their presentation that are weak and strive to improve upon them.

To provide students with rich problem-solving tasks, teachers are finding they need a variety of resource materials to supplement their textbook presentation of content areas. Teachers realize they do not have to rely ex-

Photo by Sarah Meyers

Vermont trains its teachers to serve as assessors. Each year they will evaluate a sample of portfolios from around the state against seven criteria to prepare a detailed analysis of student progress.

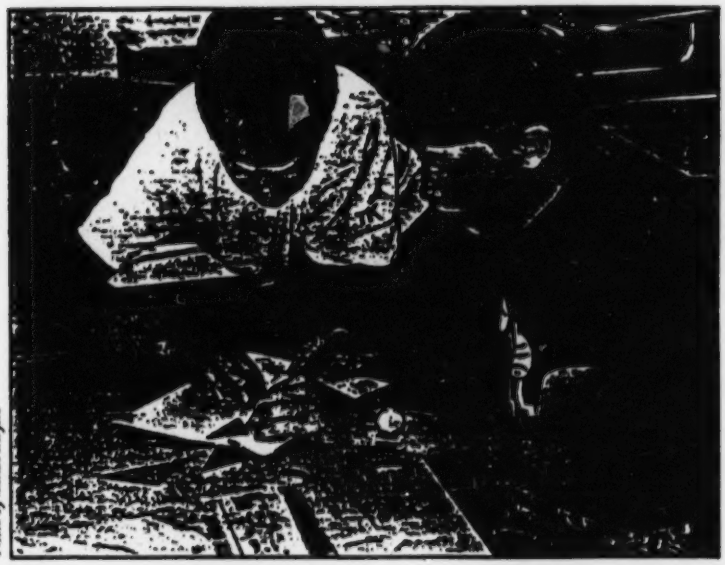


Photo by Sarah Meyers

Students in fifth grade are calculating fractions using tangram pieces. Students are asked to defend their decisions.

> "Administrators always need to be involved with their teachers and sensitive to their needs..."

clusively on a textbook but can use the text as one of many resources.

Problem solving cannot be an add-on to the curriculum. It needs to be an integral part of instruction. Teachers need to change the emphasis of instruction from computation and one-step word problems to complex mathematical tasks that engage students, tasks for which the student may not have a readily available strategy. Teachers are finding that introducing concepts through manipulatives and group work allows students to investigate mathematical ideas more thoroughly than memorizing rules given to them by the teacher.

Powerful Tool

Teachers quickly realize just how powerful a tool the student's portfolio is. It presents a wonderful profile of the student's evolution as a mathematician. It also provides an evaluation of various aspects of the teacher's program, including instructional opportunities, content areas, and student empowerment.

First, the portfolio of student work should show that the student was given a variety of instructional opportunities. These opportunities should include integrating problem solving and technology, interdisciplinary work, group work, use of manipulatives, and real-world applications. Teachers must acquire skills needed to present these opportunities in their classrooms.

A second aspect the portfolio will reveal is that problem solving was not merely an add-on to the curriculum, a two-week lesson at some point during the year. Problem solving and investigations need to be integrated into all the content areas, and the student's portfolio should demonstrate this integration.

Third, the development of mathematical power will require teachers to provide a classroom environment that encourages problem solving and models the curiosity, flexibility, and reflection we must instill in students. As students confront problems they have no ready means to solve, they begin to exhibit their power as mathematicians. They become risk-takers who persevere through challenging and engaging tasks. They prove they are curious about mathematics, flexible in approaching problems, and even able to reflect about their growth as mathematicians.

Having one's program evaluated through portfolio assessment—that is, in terms of the progress made by one's students on complex tasks—can see[m] threatening. As teachers grasp t[he] ramifications of this new assessment program, they realize administrators have a crucial role to play in providing strong and positive support.

Crucial Support

Administrators always need to be involved with their teachers and sensitive to their needs, but this supportive engagement is especially important when major instructional changes are being implemented. The administrator will be asked to support the teacher in informing school board members, parents, and the community about the portfolio assessment process and the results that are generated.

Educating community groups that are interested in and responsible for education must parallel the teachers' experiences in the classroom. If portfolio assessment is to deliver on its great potential, educating parents and community leaders fully about its elements and its utility is as important as training teachers.

A week-long training institute m[ay] not give a teacher the confidence needed to understand a new philosophy of assessment, let alone make meaningful changes in instruction. The teacher new to portfolio assessment will need a network of colleagues who can meet regularly to share ideas, materials, frustrations, successes, anxieties, and triumphs. This exchange needs to go on among the teachers in any school where portfolio assessment is being adopted. It also may mean getting help from outside sources.

In Vermont, educational planners divided the state into 17 networks, each with a writing and mathematics network leader. These leaders are teachers interested and involved in the portfolio assessment process. They are responsible for contacting each school in their area regularly to learn what services the school feels are necessary to help their teachers adopt the portfolio assessment process. A network coordinator trains th[e] network leaders in using supplemen[-] tary materials. The leaders then make presentations to schools to help them understand the criteria and to help

teachers change the emphasis of their instruction to meet the criteria.

Once the teacher understands the criteria and has provided students with rich problem-solving opportunities, the actual assessment takes place. The teacher must be trained to recognize how well each criterion is being met. The training includes having available anchor or benchmark pieces of student work to indicate each level of attainment for each criterion.

There is no substitute for experience. A teacher needs to see a range of student work to distinguish levels of attainment. Teachers must check their reliability in scoring with other teachers to be sure they see the standards through a common lens. In Vermont, this training is done in the networks. Each network checks its reliability within the network and then checks reliability among networks.

This is an exciting time in education. For too long, what has been taught in the classroom has been driven by what is easy and quick to assess. Now all those involved in education are discussing what it is they value in educating and how best to assess what they value.

Any changes in assessment, however, will call on educators to develop a new and richer understanding of assessment techniques—and almost certainly to make substantive changes in instructional methods as well. Teachers are trying to come to grips with new assessment tools and the implications they bring for classroom instruction. The success of any alternative system will depend on the skills the teachers are enabled to develop—and the support they receive.

Clare Forseth serves as a statewide trainer in portfolio assessment for the Vermont Department of Education.

RESEARCH REPORT

Reprinted, with permission, from The Executive Educator (February). Copyright (1993), the National School Boards Association. All rights reserved.

PORTFOLIO ASSESSMENT

By Susan Black

If your school is considering portfolio assessment as an alternative to traditional tests, take note: The new approach has problems of its own

ALONG WITH BOOK BAGS AND LUNCH BOXES, many students now tote something new to school—portfolios of their work. The use of portfolios is becoming increasingly popular in U.S. schools as teachers look for alternatives to traditional tests to measure student progress. But so new is the portfolio concept that there isn't yet much research to guide educators in setting up new systems. And that should signal a go-slow approach.

In fact, one recent research report from the RAND Corp., evaluating Vermont's portfolio assessment program, points to some serious problems with portfolios that should serve, according to report author Daniel Koretz, as a "warning call for people to be a little more cautious."

Portfolios appeal to educators for good reason, though. Researchers have long noted that traditional tests, standardized or otherwise, have clear limitations. According to Joan L. Herman and S. Golan, for example, the content of the tests too often determines what is taught. These researchers (along with many others) find that tests "narrow the curriculum" to basic skills rather than higher-order thinking skills. Another researcher, Lorrie Shepard, reports that even when students do well on tests (often because teachers "teach to the test"), it doesn't mean they've learned anything valuable. Most likely, Shepard says, what they've learned is to take tests well.

Many teachers, for their part, maintain there's little match between what they teach and what tests measure. Teachers who stretch children's minds beyond simple memorization and who emphasize group problem solving and cooperative learning protest that standardized tests don't reflect their students' real knowledge and abilities.

Portfolios are one answer in the search for alternative ways to assess students' performance. Portfolios are supposed to represent what students know—to show, over time and in a variety of ways, the depth, breadth, and development of student's abilities, according to researchers Lorraine Valdez Pierce and J. Michael O'Malley.

But as any teacher who's tried portfolios will attest, deciding to use them is the easy part; it's much harder to ensure portfolios accurately record and measure student performance. Some state education departments and research labs do offer workshops and booklets on portfolio assessment. But teachers usually find it's best to experiment and develop their own strategies to fit their subject areas and their classrooms. The work of Ohio State University researcher Robert Tierney and his public school teacher colleagues encourages teachers along those lines.

But even under supportive circumstances, the change from traditional testing to portfolios isn't always easy. Judith Arter, a researcher with the Northwest Regional Educational Laboratory, says few teachers who are excited about the possibilities of using portfolios have worked out exactly what they mean by portfolios or how they should be used. And, says Arter, most teachers haven't anticipated or addressed the fallout issues that can accompany portfolio assessments.

Using portfolios without a clear plan can lead to misunderstandings with parents, administrators, and students. Teachers might also find some tasks bewildering, such as setting acceptable standards for student work, coordinating assessments with grading requirements, and storing archives. If teachers feel overwhelmed with the planning details, they might forsake their good in-

Susan Black is an education consultant who lives in Penn Yan, N.Y.

tentions. And if that happens, student portfolios might end up in the circular file.

The Vermont experience

In one of the best known portfolio programs, Vermont teachers helped the state department of education design a statewide system of portfolio assessment as one way of evaluating the results of a new writing program. The portfolios are now used in grades four and eight, with grade 11 to be added soon. A typical Vermont fourth-grader's portfolio contains these pieces: (1) a table of contents listing the pieces the student has selected; (2) the best piece of writing as chosen by the student; (3) a letter—written by the student to the teacher and other reviewers—about the best piece, explaining why the student chose it and the process used to produce the final draft; (4) a poem, short story, play, or personal narrative; (5) a personal response to an event, program, or item of interest; and (6) a prose piece from any subject area other than English or language arts.

Vermont, which also uses portfolios to evaluate students' math progress, is a pioneer in implementing the approach, beginning in 1988. But the state is also a pioneer in coming face-to-face with the problems associated with portfolios. The evaluation by the RAND Corp., released in December, showed rater reliability—that is, the odds that two different teachers would rate a portfolio the same way—to be very low. The recommendations put forth in the report include improving training for teachers in how to score portfolios accurately and making changes in the scoring system itself, which the report suggests might be too complex. The report's author, Daniel Koretz, says Vermont is actively looking at instituting some of the changes and that RAND will continue its evaluation.

Koretz, a resident scholar at RAND, says those who begin using portfolios often want to accomplish two things: Improve what goes on in the classroom, and find a good assessment tool. But, he says, those two goals can be at odds with each other. Improving what goes on in the classroom often means broadly training every teacher in rating student work. Assessing students accurately, on the other hand, could well require training only a small number of teachers at a time, but training them carefully and thoroughly.

The question that needs to be addressed, Koretz says, "is how to compromise between a powerful educational intervention and a decent assessment program."

Koretz also says it's important for school districts that opt for portfolio assessment to put in place a means for assessing the program. "People ought to have realistic expectations about how quickly [implementing a portfolio system] can be done and how it will come out the first time," he says. "A lot of people around the country have unrealistic expectations."

Making way for portfolios

Before moving toward the use of portfolios, other researchers agree, teachers must first think through their reasons for using this alternative assessment approach. Do they want to improve curriculum and teaching, or to assess student work—or perhaps both? (See the box on this page for a list of possible purposes for portfolios.)

Teachers also must define exactly what they mean by portfolios. Judith Arter defines a portfolio as "a purposeful collection of student work that exhibits to the student and others effort, progress, or achievement in a given area or areas." F. Leon Paulson and Pearl R. Paulson emphasize process over product in their definition: "A portfolio is a carefully crafted portrait of what someone knows or can do." Teachers might find someone else's definition suitable, or they might choose to write their own.

Then there are the nitty-gritty decisions. What should be included in a portfolio? Who should select the contents? Should a portfolio reflect only a student's best work, or should it represent a spectrum of accomplishments and efforts? Should bulky items (such as science

WHY USE PORTFOLIOS?

Here are some reasons to use portfolios, as set down by English and language arts teachers in California. These reasons are excerpted from *Testing for Learning*, by Ruth Mitchell, and adapted from *Portfolio News*, published by Portfolio Assessment Clearinghouse, c/o San Dieguito Union High School District, Encinitas, Calif.

1. As a teaching tool
- —to provide students ownership, motivation; a sense of accomplishment, and participation.
- —to involve students in a process of self-evaluation
- —to help students and teachers set goals
- —to build in time for reflection about students' accomplishments
- —to aid in parent conferences

2. Professional development of teachers
- —to study curriculum and effective teaching practices
- —to allow for better staff communication
- —to reduce the paper load
- —to identify school strengths and needs for improvement
- —to build a sequence in writing instruction

3. Assessment
- —to serve as an alternative to standardized testing
- —to serve as a college application and high school placement vehicle
- —to replace competency exams
- —to serve as a grade or end-of-year culminating activity
- —to provide program evaluation
- —to supplement or substitute for state assessment tests

4. Research
- —to examine growth over time and progress in students' writing
- —to look at the revision process.

SELECTED REFERENCES

Alexander, L., et. al. "The Nation's Report Card: Improving the Assessment of Student Achievement." Cambridge, Mass.: National Academy of Education, 1987.

Althouse, S. M. "A Pilot Project Using Portfolios to Document Progress in the School Program." Paper presented at the annual meeting of the International Reading Association, May 1991.

Arter, J. "Curriculum-Referenced Test Development Workshop Theory: Using Portfolios in Instruction and Assessment." Portland, Ore.: Northwest Regional Educational Laboratory Nov. 1990. ERIC Document No. ED 335 364.

Ballard, L. "Portfolios and Self-Assessment." *English Journal*, Feb. 1992, 81, 46-48.

Cooper, W. and Brown, B.J. "Using Portfolios to Empower Student Writers." *English Journal*, Feb. 1992, 81, 40-45.

Herbert, E.A. "Portfolios Invite Reflection—from Students and Staff." *Educational Leadership*, May 1992, 58-61.

Herman, J.L. "What Research Tells Us About Good Assessment." *Educational Leadership*, May 1992, 74-78.

Herter, R.J. "Writing Portfolios: Alternatives to Testing." *English Journal*, Jan. 1991, 90-91.

Hetterscheidt, J., et. al. "Using the Computer as a Reading Portfolio." *Educational Leadership*, May 1992, 73.

Knight, P. "How I Use Portfolios in Mathematics." *Educational Leadership*, May 1992, 71-72.

Koretz. D., et al. *The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program, Interim Report*. Washington, D.C.: RAND Institute on Education and Training, Dec. 1992.

Mitchell, R. *Testing for Learning*. New York: The Free Press (Macmillan, Inc.), 1992.

New York State United Teachers. "Multiple Choices: Reforming Student Testing in New York State. A Report of the NYSUT Task Force on Student Assessment." Jan. 1991.

Paulson, F.L., and Paulson, P.R. "The Ins and Outs of Using Portfolios to Assess Performance (Revised.)" Expanded paper presented at the joint annual meeting of the National Council of Measurement in Education, April 1991.

Paulson, F.L., and Paulson P.R. "The Making of a Portfolio." Pre-publication Draft, Feb. 1991, 1-11.

Shepard, L. "Will National Tests Improve Student Learning?" *Phi Delta Kappan*, Nov. 1991, 232-238.

Tierney, R.J., Carter, M.A., and Desai, L.E. *Portfolio Assessment in the Reading-Writing Classroom*. Norwood, Mass.: Christopher-Gordon Publishers, Inc., 1991.

Valencia, S.W. "Alternative Assessment: Separating the Wheat from the Chaff." *The Reading Teacher*, 1990. 43, 60-61.

Wiggins, G. "The Case for Authentic Assessment." ERIC Clearinghouse on Tests, Measurement, and Evaluation, Dec. 1990.

Yunginger-Gehman, J. "A Pilot Project for Portfolio Assessment in a Chapter 1 Program." Paper presented at the Annual Meeting of the International Reading Association, May 1991.

projects) be considered? What should the school keep for its permanent records? How should teachers communicate students' achievement to parents? How should teachers evaluate portfolios? What other kinds of assessment should be used, if any?

Answers to some of these questions can be found in research. Judith Arter notes that portfolio contents should be chosen according to their purpose. She finds that, in general, teachers either require certain items from each student or, with students, choose samples of work that reflect growth and development in a specific subject area.

F. Leon Paulson and Pearl R. Paulson find portfolios should be more than just collections of students' work. Portfolios, they say, ought to include students' narratives about how they produced the contents and about what they learned. Students' written reflections about their learning might be among the most valuable pieces in the portfolios, these researchers say.

The Paulsons maintain that "students own their portfolios," so they—not teachers—should create their collections and review their selections. They suggest that portfolios should tell a student's story, and anything that helps tell that story could be included—classroom assignments, finished or rough drafts, work students develop specifically for the portfolio to show their interests and abilities, self-reflections, and observations and comments by teachers or parents.

In her report, Sharon Althouse provides diagrams to guide teachers, students, and parents as they choose items for students' reading and writing portfolios. (Althouse's study found that teachers and students most often selected *writing* samples for portfolios.)

California math teacher Pam Knight encourages her algebra students to choose from their semester's worth of classwork and homework to construct well-balanced portfolios. Knight's students are likely to include long-term projects, daily notes, journal entries about difficult test problems, scale drawings, best and worst tests, and homework samples.

In some remedial programs, students design portfolios that align with their individual education plans. In Pennsylvania's Eastern Lancaster County School District, for example, teachers first identify two or three goals for each student in Chapter 1 reading and math. (Students may add goals of their own.) Portfolios are likely to include written compositions, records of books read, examples of drafts and revisions of written work, samples of math processes and problems, and original math story problems.

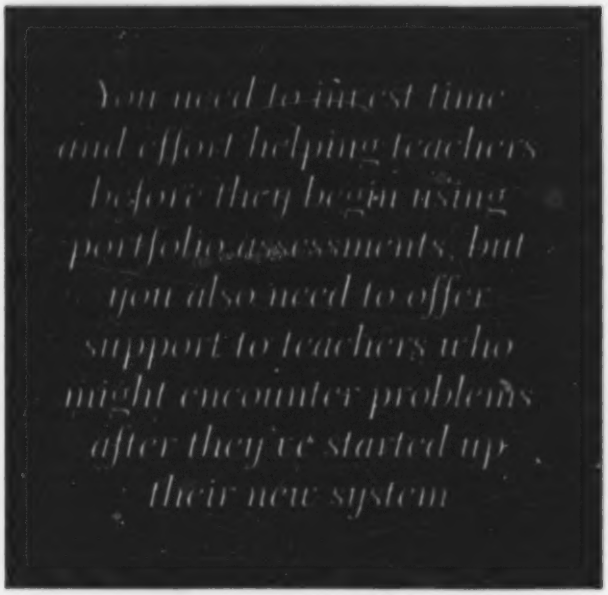
Teachers and researchers also report successful ventures using portfolios with high school science and social studies students. Missouri teachers use computer portfolios to capture their fifth-graders' reading progress. In Wyoming, elementary students use laser disk technology to record their verbal ability, physical accomplishments, artistic achievement, and self-assurance. Lorraine Valdez Pierce and J. Michael O'Malley

BEST COPY AVAILABLE

describe how elementary and middle school students learning English as a second language use portfolios to show oral language and reading skills.

Other concerns

Time and grades are among the other issues to consider before going ahead with portfolios. Managing portfolios takes time, that precious classroom commodity. But, researchers report, teachers who change from traditional assessment to portfolio assessment are more likely to manage their time without frustration if they change teaching styles at the same time. Rather than continuously assigning and grading workbook lessons, teachers should prompt students to learn through writing and exchanging ideas. Teachers can more efficiently and effectively guide instruction through cooperative learning groups. And teachers should hold conferences with their students to reinforce and motivate their learning and, when necessary, to reteach prerequisite skills.


You need to invest time and effort helping teachers before they begin using portfolio assessments, but you also need to offer support to teachers who might encounter problems after they've started up their new system

Then there's the sticky issue of grades. How can teachers assign unit grades or report card grades (usually required by the district office, and perhaps by the state education department) when they're assessing students' portfolios for effort, progress, and insight as well for as specific achievement? Some districts are experimenting with new kinds of report cards—using checklists and narratives, for example—that more closely reflect their new assessments.

It's imperative that teachers inform and educate students and parents about new grading systems. Even when they're pleased with their portfolios and their teachers' comments, some students demand familiar letter grades—especially when they're accustomed to earning A's and B's. Sometimes parents give portfolio assessment systems cool receptions because they don't understand the new evaluation reports. They might prefer their children's report cards to look just like the ones they used to bring home from school.

It's important to discuss a new portfolio assessment system with students and parents. In Sharon Althouse's Pennsylvania pilot project, students doubted their parents would understand the portfolios—even when teachers enclosed letters explaining the new assessment plan. And, Althouse reports, students often changed the contents of their portfolios when they knew their parents would examine their selections. Though parents expressed appreciation and approval of the portfolio system, they provided skimpy answers or no answers at all to a short survey about the new method.

Finally, other considerations might arise. In high schools, students and parents might object to portfolio assessment on the grounds that college admissions offices require grades and class rankings. And at any grade level, serious questions remain about the objectivity of portfolio assessment. Any program of portfolio assessment must address the possibility that assessments might be biased on the basis of race, sex, or cultural orientation or overly generous so as to bolster students' self-esteem.

A successful start

If your school or school district is considering using portfolio assessments, what can you do to help the new approach succeed? You can begin by setting agendas for staff planning and staff development. The Northwest Regional Educational Laboratory (NWREL) proposes tackling these topics: purpose, curriculum and instruction, content, assessment, management and logistics, and staff development. Specific questions to discuss, NWREL researchers say, include: What are the purposes for using portfolios? How will portfolios reflect the school's curriculum? How must instruction change to support portfolio assessments? What is acceptable in a student portfolio? Who owns the portfolio, and who chooses the contents? What other types of student assessment will the school use? How will portfolio assessments be coordinated among grade levels? How will assessments be communicated to parents?

You need to invest time and effort helping teachers *before* they begin using portfolio assessments, but you also need to offer support to teachers who might encounter problems *after* they've started up their new system. Peer coaching—pairing teachers who are reluctant or uneasy about using student portfolios with teachers who easily are incorporating the new method in their classrooms—might help teachers who want to throw in the towel and return to traditional testing alone. Workshops at which teachers examine models and then work out their own plans can also help get portfolio assessments off to a smooth start.

It's a long and often rocky road to institutionalize any innovation in education. Even when teachers are eager to accept a new plan, you can bet the process will be one of fits and starts. When it comes to developing portfolio assessments, you'll need to encourage teachers not to give up when they face difficult issues. But you'll also have to remind them—and yourself—to go slow with this new approach to assessment. ■

Reprinted with permission from T.H.E. Journal, September 1992.

Assessing the Outcomes of Computer-Based Instruction: The Experience of Maryland

by DR. GITA Z. WILDER, Research Scientist
and MARY FOWLES, Senior Examiner
Educational Testing Service
Princeton, N.J.

As computers have been introduced more widely into schools, teachers and school administrators have been called upon increasingly to justify the expenditures for hardware and software. Typically, school boards and taxpayers have demanded evidence that computers are making a difference, usually in student achievement and most often measured by scores on standardized tests.

Such demands take little account of the actual conditions that mark the implementation of technology in schools. These include the fact that implementation is a process that proceeds over a period of several years; that computers are used in a wide variety of ways by teachers and schools and for a range of purposes; and that the widespread use of instructional computing is too new to have been supported by a systematic body of research about what does and does not "work." Moreover, many schools and districts have only loosely specified objectives for the technologies that they adopt.

This paper describes one effort by several school districts to assess the effects of introducing computers into schools. The effort spanned a three-year period. At the end of the three years, all of the participants agreed their work had only just begun.

A uniform evaluation plan seemed out of the question.

■ The Sites

In 1988, the Potomac Edison Company, the utilities company that serves three counties in western Maryland, initiated a program through which the firm donated computers and educational software to schools in the areas it serves. From the outset, one key project goal was to demonstrate the potential benefits of technology for education.

Potomac Edison provides hardware (networked computers, typically), software and training; the schools develop a plan for using the technology and provide release time for teachers to participate in training.

Staff members Mary Fowles and Gita Wilder of Educational Testing Service (ETS) worked with Barbara Reeves and Patricia Mullinex of the Maryland State Department of Education (MSDE), officials of Potomac Edison and with members of an Evaluation Committee of the larger project to develop a plan for evaluating this effort to introduce technology into the schools.

■ The Challenge

Each district (and sometimes each school within a district) developed its own plan for using the technology. Moreover, the computers were utilized by different districts and/or schools for very different instructional purposes. Thus, a uniform evaluation plan seemed out of the question.

As a result, the first year of the project was spent in a series of planning sessions that involved school and district administrators in addition to representatives of Potomac Edison and MSDE. The group agreed the task of evaluating so diffuse a "treatment" was beyond the reach of any conventional program-evaluation design. At the same time, it was also acknowledged that some form of accountability was called for, and that standardized test results, although judged as inappropriate, would have to be the default measure should the group fail to devise its own approach to evaluation.

There were, moreover, several levels of "outcome," each with its own set of questions about the effects of computer use on individual students, on teachers and classrooms, and on schools. The broadest forms of the questions were, how does the introduction of computers into schools affect the learning process for students, the instructional process of teachers, the classroom process for both students and teachers, and the organizational processes of schools?

■ The Evaluation Process

The group reviewed (and agreed to incorporate) a set of questionnaires developed by the MSDE for use in a statewide evaluation of computer use. At the same time, there was a strongly felt need to collect data that reflected the specific activities and goals of the Potomac Edison project. By the end of the first year, it had become clear that the teachers who were using the technology were the most appropriate participants in any effort to document the accomplishments of the project.

The New Assessment Measures Committee of the Maryland Education Project first met in winter 1990 to devise an overall strategy for data collection. Because most of the computers had been installed in elementary schools, the committee was made up mainly of elementary school teachers, administrators and computer coordinators.

■ Work Samples as a First Step

The first meeting of the group was devoted to defining the task and to general discussion of performance assessment as a potentially fruitful approach to this new area of student progress. Members of the committee agreed, at the close of the first meeting, to devote the spring semester to collecting samples of their students' work that appeared to demonstrate the unique contributions of the computer to the learning process. Each teacher identified a set of computer-based assignments that might yield appropriate work samples from his or her class. Although general guidelines for the collection of such samples were provided, the process was purposely left somewhat open-ended so that particularly innovative and creative work could be included even if it failed to conform to the guidelines.

In organizing their samples for presentation, committee members were instructed to provide a context for each sample:

- the configuration of hardware and software with which each sample had been generated;
- the class and subject area from which the sample came;
- the assignment that resulted in the work sampled;
- the ability level(s) of the students whose work had been sampled;
- what the work sample demonstrated about the student's learning; and
- the feasibility of this as a model or prototype assignment for other classes.

At the second meeting of the committee, participating teachers brought work samples and supplied information about the contexts in which they had been generated. What was immediately apparent was the range of circumstances within which the teachers worked (from labs of 15 or more computers through single computers shared by four teachers), and the variety of solutions they had applied to their particular circumstances.

The work samples were a rich source of inspiration. They ranged from "books" of writings produced by individual first-grade students through documentation and the outcome of a class project to assign addresses to all of the houses in a small town that had previously lacked a system of addresses.

The most common application among the committee members was word processing; numerous examples of student work were provided that illustrated and explained how students engaged in the writing process. The second most common application was the organization and presentation of information from diverse sources, often facilitating complex problem-solving that could not have taken place without the computer. A third was math drill-and-practice as a way to support classroom instruction.

In the lengthy discussion that followed presentation and review of the samples, teachers talked about the ways in which the technology had enhanced teaching and learning. This discussion revealed that the benefits teachers perceived from their own and students' involvement with the computers clustered into a limited number of categories. These categories, called "domains" for purposes of this study, were further discussed and refined by the committee in a subsequent meeting.

■ Identification of Domains

In their discussion of the effects of the use of technology in their classrooms, the teachers were instructed to focus on areas in which the computer appeared to be making a real difference in instruction, rather than offering an alternative medium for "usual" classroom activities. The distinction is best illustrated by using the writing process as an example.

Many teachers reported leading their students through the writing process. Most agreed that the unique contributions of the computer to the writing process revolved around the capacity of the computer for making revisions easy and visible, and for creating products that increased students' pride in

Particularly creative work could be included even if it failed to conform to guidelines.

their products and interest in reading the writing of others. Because of the computer, the quality of the students' writing improved, more people were interested in reading it, and communication became more effective.

From the general discussion of the domains emerged a more specific list on which to focus data collection activities of the coming year. The domains so identified included:

- use of the writing process (referring to the many ways in which the computer illuminates the writing process and makes it easier for students to write and become both proficient and prolific in writing);
- production and performance for purposes of communication;
- integration of reading and writing;
- gathering, integrating, analyzing and using information;
- solving complex problems;
- adapting instruction to the needs of individual students;
- motivation (to engage in learning);
- efficiency, productivity and quality of the teaching/learning process;
- application/generalization (referring to spontaneous transfer of learning from one area to another on the part of students); and
- comfort and competence with technology.

These domains were not mutually exclusive, nor were they intended to be. Their function was to guide collection of work samples that would offer evidence of the domains in question. Any given sample might offer evidence of several domains at once.

A set of matrixes describe and provide examples of graded performance using technology.

■ Framework

The group then devised a plan for collecting work samples more systematically during the 1991 school year. Each committee member selected specific domains in which to concentrate data collection activities, identified one or more work samples collected as evidence of learning in the designated domain(s), and developed a unique plan of data collection.

From the work samples, a framework for assessing student progress had to be created. Following a (school) year of data collection, the committee re-assembled to consider the work samples and re-consider the domains. Although some domains proved more fruitful than others in the ease with which appropriate work samples could be provided, the range and variety of samples was most impressive.

Working in small groups, each of which focused on a sub-set of the domains, committee members reviewed work samples from the perspective of the evidence each provided about the level of competence demonstrated by students (and often teachers) in the domain in question.

The results of this process were a set of scaled indicators of performance within specific domains that might be used to assess progress in the application and use of technology. Teacher and student progress were described separately. And, by agreement of the committee, each set of indicators included a "top performance" rating that transcends the highest level demonstrated in any of the work samples. This rating reflects the committee's recognition that new technology and new applications thereof proceed at a rate that cannot possibly be anticipated. The transcendent rating allows for the incorporation of unforeseen, creative uses of the technology.

The ratings exist as a set of matrixes that describe and provide examples of graded performance using technology within a specified set of domains. The committee agreed that development of the matrixes is only the first step in an active process that will involve others outside of the committee. Committee members plan to apply the matrixes in their own classrooms, schools and districts. Staff members from the MSDE are disseminating the matrixes and inviting reaction from a wider set of schools and districts within the state. And readers are encouraged to contact the authors, or the Maryland Education Project at the address below, to obtain copies for their own use.

Undoubtedly, the matrixes will be revised with time and use. Yet all agreed that they, and the process that produced them, were invaluable alternatives to the use of standardized test results to assess the effects of implementing technology in schools. ■

For more information or copies of the matrixes:
Educational Testing Service
Attn: G. Wilder or M. Fowles
Rosedale Road
Princeton, NJ 08541-0001

Gita Wilder *is a research scientist in the Division of Education Policy Research at Educational Testing Service and director of a program of research on the organizational context for change. The focus of that research is school change engendered by the implementation of technology.*

Mary Fowles *is a senior examiner in the Humanities Group, Test Development Division, for School and Higher Education Programs at Educational Testing Service. Her work includes the development of assessment measures in literature and writing. Most recently, she has collaborated with several school systems and states in the development of portfolio assessments for use in a variety of subject areas.*

Reprinted with permission from Elliot W. Eisner, Stanford University, CA.

Why Standards May Not Improve Schools

Elliot W. Eisner

If we value student work that displays ingenuity and complexity, we must look beyond "standards" for evidence of achievement.

Few ideas are more central to the educational reform movement currently underway in America than that of standards. Virtually everyone thinks we need them and that efforts to improve our schools will certainly fail unless standards can be developed and made public. Standards will provide both targets and incentives. Students will know what to aim for; teachers will understand how effective they have been; and, perhaps most important, the public will know how well schools are succeeding.

Amidst the enthusiasm, nevertheless, an undercurrent suggests that standards may not be the answer and that the concept itself may impede the realization of the vision of education that many educators hold.

At the outset, we should acknowledge that the term "standards" has multiple meanings. Standards can refer to those targets at which one aims. Standards can serve as icons against which student performance is compared. The term "standard" can also refer to something that is common or typical. Standards are sometimes used in relation to a rite of passage that provides access to future opportunities. For example, "He has met the standards we have set and is now able to practice medicine."

In addition to these conceptions, standards, as Dewey pointed out in *Art as Experience*, are units of measurement.[1] We have a standard unit of measurement we call ounces, another we call pounds, and still others called inches, feet, miles, kilometers, and statute miles. All of these standards are socially defined units that quantify qualities. What we experience as height, we can transform into feet and inches. What we experience as heat, we can transform into temperature.

There is, of course, a fundamental difference between the experience of heat or height and their description through standards we call centigrade or inches. Knowing heat as an experience is not like knowing heat as temperature. Thus, one important feature of a standard as a unit of measurement is that it functions as a symbol that possesses none of the qualities of what it has measured.

We seek work that displays ingenuity, complexity, and the student's personal signature.

Another feature of standards, and this feature is crucial from an educational perspective, is that, as a unit of measurement, a standard is a vehicle for describing, rather than appraising, a set of qualities. When we apply standards in Dewey's terms, we get answers to questions that pertain to matters of amount. Yet what we want regarding the outcomes of our teaching are not simply matters of amount but matters of goodness. We want to achieve, or help students achieve, what we or they value.

Given this latter conception, standards by themselves will never be adequate for determining whether what a student has done or understands is of value. To determine matters of value, we need something more.

Criteria, Not Standards

The something more we need are *criteria*. Criteria facilitate the search for qualities we value within an essay, a scientific experiment, a painting, a work of history, and the like. These works, Dewey argues and I concur, are not susceptible to measurement by standards, although they are amenable to appraisal by criteria.

Having said that, it should be acknowledged that once a standard is assigned a value, it *can* be applied to those forms of performance or products for which there is a fixed correct answer. With a standard, we simply lay down the appropriate unit and count the correctness or incorrectness of the responses. We can apply standards to spelling and to arithmetic. We can apply standards to most forms of punctuation and to the determination of standard grammatical usage. While the application of standards to such topics is useful and efficient, standards do not represent the most important ends we seek in education. What we value in education is not simply teaching children to replicate known answers or to mimic conventional forms. We seek work that displays ingenuity, complexity, and the student's personal signature. In short, we seek work that displays the student's intelligent judgment. Work of this kind requires that we also exer-

cise judgment in appraising its value.

This judgment depends, I am arguing, on the availability of criteria. The application of criteria requires the exercise of judgment, the ability to provide reasons for the judgments we make, and an understanding of which criteria are relevant to the genre of the work. Applying criteria is a much more complicated and intellectual enterprise than applying a standard. Hence, when we talk about using standards as a lever for educational reform, we grossly oversimplify what is required.

The problem of assessment, however, is considerably more complicated than the distinction between standards and criteria that I have just drawn. Within the context of schooling, teachers do—and indeed they ought to—take into account not only the qualities of the student's work but other considerations that pertain to the individual student. Experienced and skilled teachers know that when they appraise a student's work, they need to consider where that student started, the amount of practice and effort expended, the student's age and developmental level, and the extent to which his or her current work displays progress. Although such considerations are not particularly relevant for appraising the work of professionals, they are relevant for appraising the educational development of the student.

But even if we were to consider only the student's work, we would still have the problem of determining how standards—even when they are relevant to the work—should be derived. Should standards be derived from the performance of the particular population of students from which a student comes or from the school district as a whole, the state, the region of the country, or the nation? Should standards be formulated from the performance of students in suburbs or inner cities? From those in the 50th percentile or in the 75th? Is it fair to have the same standards for all students when we know that some students come to school without breakfast while others are driven to school in expensive automobiles? Should we employ the same standards in schools that barely have enough resources to open their doors while others are the educational equivalent of Neiman Marcus? Just what is appropriate, and just what is fair?

I do not value schools that regard children as an army marching toward fixed and uniform goals.

Furthermore, what do we do with the information after the standards have been applied? Will we know what needs to be done by examining the performance of students in relation to fixed standards?

It is characteristic of our culture, and particularly our attitude towards education, to seek some magic wand, some golden lever, that we can employ to make things right. I do not argue here that our schools are in good shape. Some schools are wonderful, many schools are dreadful, and most schools, from my perspective, need to generate a much greater sense of intellectual vitality and challenge than they now possess. I cannot help but wonder whether this emphasis on standards is likely to move schools in the direction that I value. I do not value schools that regard children as an army marching toward fixed and uniform goals. Standardization is already too pervasive in our culture. We need to celebrate diversity and to cultivate the idiosyncratic aptitudes our students possess. Certainly, an array of common learning is appropriate for almost all students in our schools, but the preoccupation with uniform standards, common national goals, curriculums, achievement tests, and report cards rings in a theme that gives me pause. I believe we would do far better to pay more attention to the quality of our workplace and to the character of our teaching than to display such preoccupations with standards. If we can create schools that excite both teachers and students and provide the conditions that improve the quality of teaching, we will do much to create schools that genuinely educate.

The current emphasis on standards will provide no panacea in education. Paying close attention to how we teach and building institutions that make it possible for teachers to continue to grow as professionals may be much more effective educationally than trying to determine through standard means whether or not our students measure up. ■

[1]J. Dewey, (1934), *Art as Experience*. (New York: Minton Balch and Company).

Elliot W. Eisner is Professor of Education and Art, School of Education, Stanford University, Stanford, CA 94305-3096.

Reprinted with permission from PHI DELTA KAPPAN, December 1991.

Assessing Alternative Assessment

In looking closely at what is happening in Rhode Island, Mr. Maeroff notes, it is possible to get a preview of the satisfactions and frustrations that may spread as the movement toward large-scale alternative assessment mushrooms.

By Gene I. Maeroff

WILLIAM, a tiny third-grader in a red sweatshirt, was sitting in an adult-sized chair, balancing on the edge of the seat as he listened intently to the woman across the table telling him that she was going to ask him about one of the several books that he had read recently. "If you know the questions I'm going to ask, it may help you decide which book you want to discuss," she said, going on to list the subjects of the questions. "I will want to know about the main part of the story, the main characters, the setting, the ending, and the major conflict or problem."

And so it was that William, aware in advance of what he would be asked, decided to answer questions about a biography of Marco Polo that he had read not long before. What ensued between child and adult resembled an informal conversation more than the assessment that it actually was. "They go quite a bit by camel through the desert," William said of the setting as he answered the questions, one by one. "He lived in Venice, and they were going to trade some things from Italy for some better things. A war broke out, and so they decided to go to China and stayed there."

GENE I. MAEROFF is a senior fellow at the Carnegie Foundation for the Advancement of Teaching, Princeton, N.J., and author of The School-Smart Parent *(Henry Holt, 1990).*

Illustration by Kay Salem

For all its attractiveness, alternative assessment is fraught with complications and difficulties.

"Does it end in China?" William was asked.

"No," he answered. "After 17 years he decided to go home and left on a fleet of Chinese junks. When Marco Polo got back to Italy, no one believed him, so they put him in prison. He wrote a book about it, and years later Christopher Columbus read it."

This encounter was an assessment of William's reading, and the only paper and pencil in evidence were in the hands of the adult who was questioning him and taking notes on his responses. None of the questions had multiple-choice or true-or-false answers, and nothing about this reading examination was norm-referenced or nationally standardized. It was an alternative assessment, which included the individual interview about the book on Marco Polo and a review of a portfolio containing samples of William's work. Later, with another assessor, William would be assessed on other tasks, participating in a small group in which he would be asked to read and discuss a different story and to write about it.

Speed and low cost were the silver bullets that enabled norm-referenced tests to carry the day.

Alternative assessments of this sort typify a movement that is capturing the imagination of educators across the country. Good teachers have historically used such methods to monitor the progress of their students, but now these approaches are being extended beyond individual classrooms to pose a challenge to traditional ways of mass testing. However, large-scale alternative assessment is still talked about more than it is used, and William's experience in his school in South Kingstown, Rhode Island, represents one of the early efforts by a state to develop a system of alternative assessments in its elementary schools.

Other states – notably California, Connecticut, Kentucky, and Vermont – as well as some school systems and individual schools are at various stages in this gentle upheaval that proponents hope will alter the way Americans think about evaluating schoolwork. Altogether, 40 states are planning some form of alternative assessment at the state level, with writing samples as the most common alternative.[1]

Rhode Island's pilot project, which began in 1989, is still inchoate as a small number of third-grade teachers in four school districts around the state collaborate with officials from the state education department and a researcher from the Educational Testing Service (ETS). Together they are struggling to devise methods of assessing students that can provide useful information while avoiding the shortcomings associated with norm-referenced tests. In looking closely at what is happening in Rhode Island, it is possible to get a preview of the satisfactions and frustrations that may spread as the movement mushrooms.

Those pursuing change in Rhode Island are continually administering and revising the pilot assessments, striving to create a framework for instruments that eventually might be used in schools throughout the state at various grade levels. They hope to forge a synergy of instruction and assessment in which each complements the other to raise learning to new levels. This approach allows a child to know in advance – as William did – what will be asked. The goal is not to spring surprises on students and catch them unawares; unlike the usual mode of testing, this is no "gotcha" game. The students, under the tutelage of their teachers, are trained to provide evidence of their own learning.

The original charge to the pilot group in Rhode Island was to try to determine how the state's "Outcomes for Third-Graders" might be measured through the use of portfolios. This goal, which fits comfortably within the nation's alternative assessment movement, has extended beyond portfolios to include other possibilities, such as performance tasks and group interviews, and it embraces outcomes in reading, writing, speaking, listening, and mathematics.

Schools presumably are doing their job if students learn something that is deemed worth knowing. This kind of assessment does not drive the curriculum; it grows out of the curriculum and is part and parcel of the curriculum. Such a philosophy is widely accepted at the elementary and secondary levels in the performing and studio arts, in athletics, and even in vocational education. But it is not readily accepted in formal academics. Think, for instance, how long it was before test-makers asked students to produce essays in order to demonstrate their writing skills instead of giving them tests with questions about writing.

A young pianist who is asked to master Beethoven's "Für Elise" becomes proficient by practicing the piece, knowing all the while that his examination will consist of playing it. A nonswimmer who is told that the measure of her ability to swim will be the completion of one lap of the pool endeavors to swim with full knowledge of the form the assessment will take. Being able to play Beethoven or to swim a lap of the pool is presumably indicative of the ability to play some other composition of equal difficulty or to swim a similar distance on another occasion.

YET, FOR all its attractiveness, alternative assessment is fraught with complications and difficulties, not unlike having to endure life with a teenager as the price for the joys of parenthood. If they are to be used widely or even as supplements to nationally standardized,

norm-referenced tests, these new assessments will have to be done more quickly, more efficiently, and less expensively than at first seems possible.

As experimenters in Rhode Island and elsewhere are discovering, the quest for alternative forms of mass testing could remain as elusive as Don Quixote's dream. The schools seem unable even to broker the logical marriage of alternative assessment and technology; they treat the two as if they dwelt on separate planets. For example, in elementary and secondary education, how often do we hear of alternative assessment that involves the manipulation of computer models or simulations?

Furthermore, there must be standardization of some sort, as Rhode Island hopes to achieve. Otherwise, there is no way to put the findings of an assessment in context. Even in Kentucky, which by 1995 is to have the first statewide assessment system that is completely based on performance, there is anguish over how to meet the state board's mandate that there be a way of comparing Kentucky students with those in other states.

Speed and low cost were the silver bullets that enabled the norm-referenced test – with its multiple-choice responses – to conquer the world of education and hold it in thrall. Meanwhile, alternative assessment, which is not so new an idea as some people think, tends to be a time-consuming, labor-intensive, imprecise exercise in which the expense mounts as nuances are weighed and scoring is done by humans, without even the benefit of grids that fit conveniently over answer sheets studded with blacked-in boxes.

On the other hand, there is the potential for teasing considerable consensus from an approach that seems at first to be hopelessly subjective. Consider the competition in such sports as diving and figure skating, in which experts have developed scoring criteria that are so extensively accepted around the world that judges find a remarkable level of concurrence. Some of the best work in assessing writing samples mirrors this sort of agreement on relevant criteria.

While it may be possible to be systematic about alternative assessment, there are ultimately no quick and easy ways to rate large numbers of performance-based tasks or portfolios or interviews or exhibits or even essays. American proponents of alternative assessment like to cite the example of England as a country that has used alternative assessments to the exclusion of norm-referenced and multiple-choice tests. That is generally true, but until recently it has mostly meant using essays for external assessment. Only lately has England made a large-scale attempt to develop for external assessment such nonwriting tasks as science experiments. Meanwhile, portfolios containing work done over time have been and will continue to be used in England for assessment within the classroom, though not for outside comparisons of students or for purposes of accountability.

Measurement experts from around the U.S. who gathered in California in March 1991 under the auspices of the federal government's Center for Research on Evaluation, Standards, and Student Testing commiserated over the difficulty of meeting the rapidly escalating demands of policy makers for alternative assessment tools. They noted the time that it will take to establish validity and reliability, the need to train teachers in alternative assessment procedures, and the amount of time needed to administer the examinations.[2]

Nonetheless, the pioneers in Rhode Island and other locales – like those who drove their Conestoga wagons into unfamiliar territory – steel themselves and press the journey forward, intent on building an assessment system that is embedded in instruction. They want an approach that attests to a student's progress over time, something that figuratively resembles a semester-length videotape rather than a Polaroid snapshot that, like a one-shot test, captures only a single moment of a child's learning.

The cynosure for each of five subject areas in Rhode Island is a set of literacy outcomes for third-graders. One aim is to assess outcomes in reading, writing, speaking, listening, and mathematics – both as individual subjects and in ways that integrate them. The intention is that teachers will then teach in this manner.

In this approach, assessment drives instruction, and instruction drives assessment, much the way the front and rear axles impel one another in a vehicle with four-wheel drive. In essence, the assessment task is part of the instruction. This notion may be revolutionary for elementary and secondary education, but in medical schools, where clinical education is often evaluated on the basis of performance tasks that may have an instructional component even during the examination, it is accepted practice.[3] When assessment is sheared off from instruction, as is customary in the commercially produced examinations given in the public schools, the findings may not tell a great deal about what students have learned from their classroom experience.

One must wonder, therefore, about the movement for a national examination system that is hurrying forward on several fronts. A great deal is heard about various plans to create tests that would be given to students at, say, fourth, eighth, and 12th grades. Yet these discussions are curiously reticent about the curriculum on which students are to be examined. Pre-

sumably, these tests will dictate the curriculum, a policy that is anathema to proponents of alternative assessment. The concern about widespread low achievement that is fueling the movement for a national test is understandable. But wouldn't it be more straightforward to decide what ought to be taught and to teach it, instead of first administering tests and hoping that they push the curriculum in the right direction?

Perhaps circuitous routes are chosen because the challenge of unifying teaching and assessment is so daunting. Assessment that is authentically embedded in instruction is not easy to fashion. This is no Operation Desert Storm. There are no smart bombs to wipe out specific problems, nor any flanking maneuvers to avoid the tedious process of repeatedly refining the assessment instruments. There is just the slow, mundane reconnoitering to find and frame the best ways of eliciting the most useful information. One can only hope that alternative assessment is not rushed onto the battlefield of testing so hastily as to produce in its unperfected form friendly fire that harms the very children who are supposed to be the beneficiaries.

And the work doesn't stop with declaring, for example, that students will submit portfolios. What should be in the portfolios? What should students be asked about the contents of their portfolios? How can some element of standardization be lent to the process so that one student's portfolio may be compared with another's? Putting less emphasis on comparisons is fine, but at some point a child and his parents have a right to know whether the child's progress is reasonable for his or her age and experience.

THERE IS an inclination among proponents to regard alternative assessment as suitable for all purposes – diagnosis, selection, and accountability at all levels. It may be that an alternative assessment that is a marvelous indicator of an individual child's academic progress will prove fairly useless for other purposes. Americans may have to decide whether comparisons are what they seek in alternative assessment or whether they prefer to use the approach for other, more individualized purposes. Incidentally, the very idea of comparing the progress of young children is seldom embraced in Europe, where the practice is generally *not* to test students in the primary years. But America is different.

Thus, one crisp New England morning at South Road Elementary School, little William and some of his classmates played guinea pigs for the assessment developers, who had situated themselves at three separate sites in the school. It was part of a continuing process of refinement that has continued through this fall.

Grant Wiggins reminds us that the root of the word *assessment* means to "sit with" a learner and seek to be sure that a student's responses really mean what they seem to mean. He adds: "Does a correct answer mask thoughtless recall? Does a wrong answer obscure thoughtful understanding? We can know for sure by asking further questions, by seeking explanation or substantiation, by requesting a self-assessment, or by soliciting the student's response to the assessment."[4]

Next door to the conference room in which William met with his assessor, Mary Fowles of ETS, four other third-graders wearing name tags sat at a table with Susan Skawinski of the state department of education. She was also carrying out a pilot assessment, talking about a story that they had read and were going to write about. The idea at this point was not so much to assess the children's learning as to assess the assessment itself. Eventually, classroom teachers are to be the assessors of their own students. And when this happens, they will need a fully developed instrument.

This was a group interview in which Skawinski would assess each student's understanding of literature by having him or her discuss and write about what had been read. "I'm going to talk to you about something you've done, and I'm going to talk to you about reading," Skawinski said to the three girls and one boy who were gathered with her around the Formica-topped table.

"In this story we have some very important characters," she said, proceeding to list and discuss the characters in "The Rooster Who Understood Japanese," a tale about a Japanese-American family that owned a rooster whose crowing woke a neighbor each morning. Their home was on the outskirts of a city. Skawinski handed out copies of a "story map" to the youngsters and, following its scheme, mentioned the characters, the setting, the problem, the attempts to solve the problem – all of which had been briefly summarized on the story map. She took notes on a pad on her lap as the students – sometimes readily, sometimes reluctantly – entered the conversation.

The box labeled "solution" was blank on the story map, and the children were asked to write in the details of what had been done to deal with the crowing of the rooster, named Mr. Lincoln. The students were permitted to go back and look at the story if they wished. This, it turned out, was a warm-up for the real task: the writing assessment. Skawinski asked the children to think about solutions in addition to the one offered by the author and prodded them to discuss their ideas with the group.

"Tie a rope around his beak," suggested Rebecca, causing Kevin, sitting next to her, to look crestfallen. He said that he had had the same idea.

BEST COPY AVAILABLE

"Send him to rooster obedience school," suggested Alicia.

And so it went as possible solutions to the conflict were proposed and discussed. "Pick a solution you think makes the most sense and write about it," Skawinski said. "It can be your own or someone else's that we talked about." In addition, the children were asked to address two other questions about the story. The following excerpts are part of what each third-grader wrote:

> "keep mr. licon in the dark because the moring he crows in but not at night."
>
> "I would put Mr. Lincoln in the dark so he would think it was still night and wouldn't crow."
>
> "talk to him in japaneas and then he will be quiete"
>
> "Put him in a room when it's dark to make him feel that it's still dark out side."

It took so long to complete the discussion and the writing that, as she collected their papers, Skawinski told the children that there would not be time for them to talk about their favorite books, which they had been asked to bring with them to the assessment. "It was a laboriously slow process," she said afterward. "I needed more time." Such complaints are frequent in the piloting of alternative assessments because a central problem is figuring out how to accomplish the assessment in a manageable time period.

Furthermore, in this case, the assessor would still have to devote time to sorting out the notes she took during the assessment. "You can't actually mark the sheets in front of the children," Skawinski remarked after the students had left.

She noted discreetly on the pad that had been perched on her lap throughout the interview the extent to which each child seemed to bring background knowledge to the story -- ideas about living arrangements in a city, an awareness that roosters crow early in the morning, and so on. She noted their familiarity with a story map as a way of setting out the structure of the story. She also observed how long it took each child to read the story silently and what responses each one had to questions about the story -- who had original ideas, who followed the ideas of others, how well individual students could go back and cite information from the story to justify what they were saying.

Mary Fowles of ETS expects that, as the assessment is improved, there will be less note-taking by the assessor. In the pilot assessments, attention has been paid to deciding what information is most important to gather about a student's reactions and what kinds of evaluation forms lend themselves to recording this information. Theoretically, it is possible for the evaluation form to be anything from a check-off list to a sheet on which detailed quotes from a discussion with a student are recorded.

Assessing a portfolio might seem to be a more straightforward job than assessing students' performance in a group interview, but the challenge is simply of a different sort. This was illustrated in the trial assessment of mathematics portfolios that Mary Ann Snider, a testing and evaluation specialist with the state education department in Rhode Island, was conducting with students at a desk in the corridor outside their classroom.

"Your teacher tells me you've been doing several things in math that you might want to share with me," Snider said to the first of the students, Nigel, as she began reviewing his math portfolio with him after a bit of casual chatter. Nigel was eager to talk and spoke enthusiastically about his work in conjunction with a medieval theme, used concurrently by many teachers in his school as a means of teaching various subjects in an interdisciplinary fashion. There had also been a schoolwide medieval festival for which the classes had gathered en masse.

Nigel showed Snider a pencil drawing he had made of intricately linked chain mail, and they discussed what he had learned about shapes by making the drawing. "What else did you do with math in the medieval project?" Snider eventually asked. The emphasis in this part of the assessment was on finding ways in which math had been integrated into other subjects, reflecting the goal of bringing about the interdisciplinary teaching that the assessment is intended to reinforce.

Nigel displayed a word puzzle that he had devised, using vocabulary words associated with the medieval period, and a separate math word problem that he had written on a medieval theme. But the math problem he had created was so convoluted that even he could not solve it. Then he showed some work that he had done on estimation. It called for counting the number of bricks in one portion of a castle wall and estimating the number in the entire wall. Snider initiated a discussion of how the drawing of the castle might be used to study shapes and asked Nigel to identify some of the shapes.

Moving beyond the contents of the portfolio and following a written script so that the assessment would be systematic, Snider posed a series of questions, some of which were more successful than others in eliciting responses that could help in assessing Nigel's understanding of math: "What has been your favorite thing to do in math this year?" "When you learn something new in math, when do you know that you understand it?" "Have any of the new things you've learned in math helped you outside of school?"

Finally, they reached the last part of the assessment, the task that Nigel would be asked to perform. A plastic bucket filled to the brim with small, square, colored tiles was put on the desk between them, and Snider held up a bag.

"I have something in this brown paper bag, and I have a riddle so you can figure out what is in it," she said. "I will give you some clues, and you can use these tiles to figure it out. Here's the first clue: there are fewer than 10 tiles in the bag. What does that mean?"

It is easier to propose outcomes than it is to set the criteria and establish performance levels.

"Less than 10."

"Does that mean I could have 10 of them in the bag?" Snider asked.

"Yes . . . ," Nigel answered hesitantly. Then, upon a moment's reflection, he changed his answer.

The goal was for Nigel to take from the bucket and put onto the desk tiles of the same number and color as the riddle indicated were in the bag. He took nine tiles of various colors from the bucket and spread them on the desk, now fairly certain that there could not be more than nine in the bag.

"Clue 2: there are two colors in the bag." And so it went, Nigel having to modify the number or color of the tiles in front of him in response to a series of clues. When he thought he had solved the problem, Snider gave him the bag so that he could take out the tiles and see if the number and color matched those he had assembled on the desk.

"Now," Snider said, "I'd like you to write a riddle for me using the tiles, and I will try to solve it."

This was not a wholly satisfactory assessment, according to Snider, who said afterward that she wished Nigel had accumulated more math work in his portfolio to reflect studies pursued more recently than the medieval project, completed some six weeks earlier. Furthermore, she worried that the warm-up period had not been adequate and that Nigel was not sufficiently at ease during the assessment. "I saw emerging evidence of understanding," she said. "He is logical but not confident. He doesn't have enough math vocabulary, and he has trouble talking about the work he did. He is not comfortable explaining. He remembered what estimation was, but he could not say enough about it to show that he understood it."

Students ought to spend time thinking about questions like those to be asked by assessors.

SO IT IS that Snider and the others who are piloting Rhode Island's assessment are carrying out their work in selected schools in East Providence, Glocester, South Kingstown, and Newport. In each of the four participating districts, a lead teacher acts as a liaison between the project and other third-grade teachers in the district. One aim of the project is to develop a format for reporting indicators of a student's progress toward reaching the state's proposed literacy outcomes. Most experts believe that the reports should be more descriptive than just a row of numbers. But the reports *will* include numbers, and the experts are striving to determine what those numbers should represent.

This attempt to add meaning to the evaluative numbers can be appreciated by examining what happened in a place that preceded Rhode Island in the shift toward alternative assessment. When Mark Twain Elementary School in Littleton, Colorado, began moving into an assessment program that would be based on the actual performances of students, a major part of the effort was devoted to developing criteria for scoring the performances. The criteria evolved through at least 10 stages, according to Monte Moses, the school's principal. "Every time we gave the assessment, we saw some student doing something we couldn't account for on the scoring rubric," Moses said.

One part of an assignment at Mark Twain, for example, calls for fifth-graders to submit a written research report. It is scored on a descending scale of 5 to 1. A report earns a 5 when it

> clearly describes the question studied and provides strong reasons for its importance. Conclusions are clearly stated in a thoughtful manner. A variety of facts, details, and examples are given to answer the question and provide support for the answer. The writing is engaging, organized, fluid, and very readable. Sentence structure is varied, and grammar, mechanics, and spelling are consistently correct. Sources of information are noted and cited in an appropriate way.

By comparison, a 3 is awarded to a written report when the student

> briefly describes the question and has written conclusions. An answer is stated with a small amount of supporting information. The writing has a basic organization although it is not always clear and is sometimes difficult to follow. Sentence structure and mechanics are generally correct with some weaknesses and errors. References are mentioned, but without adequate detail.[3]

The bottom line is that it is easier to propose outcomes than it is to set the criteria and establish the performance levels that are represented by various achievements. Moreover, if students themselves are to take responsibility for their own work, the criteria must be spelled out in ways that are understandable to children. Then the students can go about learning how to do what is expected of them. In general, it is desirable that a student spend time thinking about questions of the type that will be asked on the assessment; those questions should promote and direct the child's learning.

In Rhode Island, some individuals administering the pilot assessments wondered after one of the pilot sessions how useful it was, for instance, to ask a student, "When you learn something new in math, when do you know that you under-

Despite England's considerable experience with alternative assessment, her efforts have been marked by controversy.

stand it?" or "What was the hardest thing about reading for you this year?" At each step of the Rhode Island project, the evaluation is being fine-tuned, as it was at Mark Twain. Questions are dropped, added, or modified.

The way that assessment can enhance the learning of students is illustrated in Rhode Island by the evaluation of speaking ability. An evaluation form serves the twin purposes of instruction and assessment. Proceeding on the assumption that the best way to get students to reach the expected outcomes is to familiarize them with the expectations, students see the criteria by which they will be rated.

In fact, the children are asked to apply these criteria to one another, using the same evaluation forms by which they will be rated. In most cases, the speech to be rated is an oral presentation of a book report. The student is supposed to be sufficiently conversant with the book to be able to deliver the report by referring to notes, not by reading a written text. One teacher found that students were somewhat reluctant to fill out the evaluation forms because they did not want to rate one another, especially if it meant saying something negative about a classmate. She encouraged them to think of the evaluation as "helping one another." The assessment consists of the following questions, each of which is answered by circling a yes or a no:

> Did the student
> • speak so that everyone could hear?
> • finish sentences?
> • seem comfortable in front of the group?
> • give a good introduction?
> • seem well-informed about the topic?
> • explain ideas clearly?
> • stay on the topic?
> • give a good conclusion?
> • use effective costumes, pictures, or other materials to make the presentation interesting?
> • give good answers to questions from the audience?[6]

After giving a report, a student has only to sort through the evaluation sheets marked by classmates to discover what he or she must do to improve. The crucial point here is that the strands of instruction and assessment are so interwoven. The Rhode Islanders are trying to help create classroom assignments that both support the growth of students in the five specified literacy areas and prepare them for assessment, as the evaluation form in speaking does. The most promising practices will be documented so that they may be shared among the teachers.

THE DIFFICULTY of banishing the glitches from alternative assessment may be sensed from a glimpse at England's attempt to develop sets of "standard assessment tasks" that are to be used nationally for the first time in conjunction with the new national curriculum. Despite England's considerable experience with alternative assessment, that nation's efforts have been marked by controversy. One teacher told of his frustration in carrying out the pilot assessment in science for 7-year-olds last spring. He said:

> Close observation of the pupils was necessary throughout. It was difficult deciding whether Dionne was counting with her fingers under the table or mentally sorting through the number bonds that she genuinely knew. During discussion about floating and sinking, it required considerable attention to assess the pupils against the 13 states of attainment.[7]

It took more than 35 minutes to test just four students on the scientific floating exercises alone, according to this teacher, and still the results were inconclusive. The proposed number of tasks to be performed in England's national assessment has been reduced by as much as two-thirds because the assessment, as originally conceived, demanded too much time. Students waiting to be assessed and those already assessed were supposed to work on their own while the teacher was off in a corner of the room assessing their classmates in small groups. But after students ended up being left unsupervised for long periods – and sometimes disrupting classmates who were being assessed – some schools sought to scrape up money to hire substitutes to oversee the classes while regular teachers were conducting assessments. There is even talk now in England of inserting some multiple-choice questions into the assessments to speed up the process.

The matter of time is also on the minds of those developing the alternatives in Rhode Island. One day, as they sat around evaluating yet another pilot assessment, the assessors wondered whether they should continue to review portfolios only in the presence of the students. "It might have been good to have had five or 10 minutes with the portfolios before the student walked into the room to discuss it," said one assessor, raising the prospect of adding even more time to a job that some believe lasts too long already.

Subsequently it was decided that each portfolio would be reviewed briefly before the student arrived. This step would not only increase the amount of time required for the assessment but would also mean that the portfolios would have to be more self-explanatory and more selectively assembled. In preparation for this change in policy, teachers were to reconsider the kinds of pieces that should go into the portfolio so as to reduce the amount of peripheral material. After all, a portfolio should not be like a kitchen gadget drawer, so chock-a-block with unrelated items that locating the corkscrew becomes a frustrating quest.

Among the sorts of writing to be weighed for inclusion in the portfolio were: 1) a favorite piece of writing of any genre, 2) a set of revisions showing the evolution of the writing process, 3) a creative and/or informative piece that responds to literature that the student has read, 4) a piece illustrating the student's understanding in a particular content area, and 5) a wild card that the student wants to include for whatever reason.

Since the content of the portfolio depends on what has been generated by assignments, the onus is on teachers to teach in ways that lead students to produce samples of appropriate work. At the same time, teachers must guide students in making proper selections from the various pieces they have written in each category, and they must prepare students to respond to an assessor in ways that will help him or her make a proper assessment. Ideally, the steps leading to an assessment and the assessment itself will be a learning experience for students.

This responsibility can be forbidding to a teacher unaccustomed to teaching in this manner. Judy Wood, a third-grade teacher in Rhode Island, who is in her 28th and probably final year of teaching, can now reflect on her participation in the development of the alternative assessment with a little less anxiety than she and her colleagues felt at the outset of the project. She said:

> At first, when they were asked to get involved, the teachers thought they would be given something that had already been completed. They were uneasy when they realized that they were getting involved in something that was only in the development stage. After a while, they understood that there was not a finished product for them and that their input would count and that there was a big, messy ingredient – the papers that were needed from the children for the portfolios. Finally, this fall, we were organized enough to know what was needed and to help the teachers in the first week or two of school so that those who were interested in joining in could start getting the folders going.

THE WAYS in which both learning and assessment might not be fully served were illustrated in one of the sessions in Rhode Island by a student who was asked to show her best work. (There has been some agonizing, incidentally, over possible differences in responses when children are asked for their "best" work as opposed to their "favorite" work.) She kept displaying examples from the beginning of the school year, although as a third-grader her work had presumably improved during the term. Difficulties arose in assessing other students when their math portfolios contained only the final step in solving a problem and showed little evidence of how that point had been reached.

In yet another instance, the assessor's task was made more difficult by what was not contained in a student's math portfolio. "We wrote some word problems with circulars from Stop 'N' Shop," said the student, Stacy, reaching into a folder and pulling out a list of questions she had written using the grocery prices in an accompanying circular published by the store. One question asked, "If you had $8.99 and bought Head & Shoulders, how much would you have left?" The reader had to find the price listed for the product in the circular and then subtract it from the amount in Stacy's problem.

Stacy was asked why all her word problems involved only subtraction. She said that there were some that called for addition, but she hadn't been able to find them. When asked about her use of math in other subjects, Stacy pulled from her portfolio pages of graphs that she had drawn comparing temperatures on various dates in Rhode Island and Hawaii, apparently reflecting work in social studies. Stacy had nothing to show when she was asked about what she had learned of division and multiplication.

During their reviews of the pilot assessments, the assessors questioned whether even the settings might have affected the assessments, because some sessions were held in large rooms where other assessments were also taking place, while others were held in smaller rooms in which only one assessment at a time took place.

Such concerns have arisen elsewhere. Researchers at England's Bristol University who studied assessments of performance tasks conducted last spring in 57 of the county's primary schools wondered about the relationship of the settings to the comparability of outcomes. In some instances, students being assessed were regularly interrupted by comments directed at them by classmates working elsewhere in the room. At some other sites, the students being assessed were the only ones in the room and had the full attention of the assessor.[8]

Despite the existence of models on which to build, there is clearly much about alternative assessment that remains problematic. There are many hurdles that teachers must learn to leap before they assume responsibility for external assessment, as indeed they must if the activity is to be financially viable.

Nor can it be taken for granted that teachers will easily shift to alternative assessment. Many teachers do not fully understand the intricacies of the kinds of examinations with which they have been working until now. Most states require no training in assessment as a part of teacher certification – and, "even when assessment training is offered to teachers, it typically fails to provide the kinds of knowledge and skills needed to produce assessment literates."[9]

There is promise in what has been happening in such endeavors as Arts PROPEL in Pittsburgh, a program in which teachers of music, of the visual arts, and of writing have been altering their teaching to weave a blend of instruction and assessment that enables both students and teachers to be reflective throughout the process of creation. But money from the Rockefeller Foundation and ex-

pertise from Harvard University and from ETS have perhaps made this venture easier to carry out than it might be in schools without such financial and human resources on which to draw.

Another matter is seldom mentioned. Those who have cited the need for equity in their rush toward alternative assessment should recognize that students who score poorly on the much-maligned norm-referenced tests with their multiple-choice responses are not necessarily going to perform better on the alternatives. In fact, there is reason to suspect that the weakest students could look even worse – though they may avoid the embarrassment of being ranked and compared on a numerical scale.

In England in the late 1980s, when the assessments that make up the General Certificate of Secondary Education were changed to put more emphasis on performance tasks (which are assessed by classroom teachers) and less on written answers, the gaps between the average scores of various ethnic groups *increased* rather than narrowed.[10] While such findings should not dampen the ardor of those eager to court alternative assessment, the suitors must nonetheless be realistic and recognize that the bride has imperfections.

Then there are issues related to the significance of being able to perform a given task. As I pointed out above, it seems reasonable to let students know that they will be examined on their ability to play a certain composition or execute a specific athletic feat. But what about when it is less clear that the performance of the task is indicative of understanding that is transferable and not merely the result of memorization that barely outlasts the assessment?

For example, New York State has used the same tasks for three years to assess the manipulative skills of fourth-graders in science. Each year a new group of fourth-graders goes to the same five testing stations as did fourth-graders the year before, where their skills are assessed in: 1) recognizing the physical properties of an object, 2) predicting, 3) inferring, 4) creating a classification system, and 5) writing a generalization. The main reason why the tasks and the equipment have stayed the same for three years – changes are scheduled for 1992 – is that making changes and training teachers to conduct new assessments are expensive.

Fourth-grade teachers throughout New York know which tasks are to be assessed, and there is nothing to prevent them from coaching their students for the performance that occurs each May. Douglas Reynolds, who oversees this science assessment for the state department of education, said that the tasks – such as knowing how and when to use a thermometer or a ruler or an equal-arm balance – are widely transferable, and, even if children practice them in advance, all they are doing is learning what they ought to know. He was quick to add, though, "If we weren't working on a shoestring and if I had an infinite pool of items and if I could go back and train people every year, then the tasks would not be the same each year."

Worries of this sort are not unique to science tests for fourth-graders. At medical schools using alternative assessments there has been concern about the impact of discussions between students who have taken the examinations and those who are waiting to do so. This worry arose in connection with a lengthy assessment that could accommodate only small numbers of students at a time, as they took turns diagnosing mock patients who simulated symptoms.[11]

Expense and time may turn out to be the brakes on the alternative assessment movement, both for the development of instruments and for their use. But thumbing through a portfolio with a student or watching a student perform a task – whatever the psychometric worth of such assessments – adds a degree of intimacy that can be refreshing in an age of depersonalized appraisal.

For instance, when the assessment in Rhode Island shifted from little William's reading to his writing, he was asked to show and discuss the story he considered his best. Asked what he found most troublesome about writing, William – the proud author of a story about a nobleman who has a feast for his friends – did not pause for a moment, issuing a response that is universal in its simplicity for authors of all ages: "The hardest part is covering the blank pages." And so it goes with alternative assessment.

1. "Performance Assessments in the States," paper prepared by the Council of Chief State School Officers for presentation to the Secretary's Commission on Achieving Necessary Skills, January 1991.
2. Robert Rothman, "Supply of New Assessment Methods Said Trailing Behind Strong Demand," *Education Week*, 20 March 1991, p. 11.
3. Vicki Kowlowitz et al., "Implementing the Objective Structured Clinical Examination in a Traditional Medical School," *Academic Medicine*, June 1991, pp. 345-47.
4. Grant Wiggins, "A True Test: Toward More Authentic and Equitable Assessment," *Phi Delta Kappan*, May 1989, p. 708.
5. "Fifth-Grade Research Performance Assessment," Mark Twain Elementary School document.
6. Excerpted from assessment materials of the Rhode Island Literacy Portfolio Project, 1991.
7. Richard Stainton, "Sinking Rather Than Floating," *Times* (London) *Educational Supplement*, 1 March 1991, p. 15.
8. Diane Hofkins, "Testing Conditions Undermined SATs," *Times* (London) *Educational Supplement*, 9 August 1991, p. 9.
9. Richard J. Stiggins, "Assessment Literacy," *Phi Delta Kappan*, March 1991, p. 535.
10. Desmond L. Nuttall and Harvey Goldstein, "The 1988 Examination Results for ILEA," paper presented to the Inner London Educational Authority Committees, March 1990.
11. Jerry A. Colliver et al., "Test Security in Examinations That Use Standardized-Patient Cases at One Medical School," *Academic Medicine*, May 1991, pp. 279-82. K

Reprinted with permission from Elsbeth Bellemere, Curriculum Director, Scarborough, Maine Public Schools.

Assessment Recordkeeping in a Non-Graded Developmentally-Based Program

ELSBETH BELLEMERE
JEANNE KING

The Scarborough School System has adopted a policy calling for 'Instructional Choices'. The policy reads, in part: "Recognizing the diversity of its students and their differing learning styles, the Board shall strive to provide a variety of programs designed to enhance learning."

To implement this policy, Scarborough has created a non-graded, developmentally-based primary level program known as the GOLD (Grouping for Optimal Learning Development) Program. As its framework and as part of its assessment process, the program relies heavily on Jean Piaget's principles of cognitive development and the structure of thinking.

Jeanne King, who teaches at the Eight Corners School, described the various components of her assessment process:

"What's important about the Piagetian framework is that it promotes matching the child to the curriculum. As a result, the assessment becomes tied to that curriculum and what kids really do. In the program, we have a three-year rotation of themes: farm, forest, ocean community. We fit the science and social studies and other concepts in where they are appropriate.

In my classroom there are children ranging in age from five to eight years old. We have no text books. We have developed sets of criteria which a student must master before he or she can move on to the next level.

In my classroom there are children ranging in age from five to eight years old. We have no text books. We have developed sets of criteria which a student must master before he or she can move on to the next level.

The cornerstone of the assessment is constant and organized recordkeeping. The teacher is not the center of attention. I'm the facilitator/observer so I'm constantly observing, taking notes, going through students' work. We save their work to check progress over time and to keep as a record of the child's accomplishments.

For example, one way I get to know students is by reading with them individually once or twice a week. When I do this, I note the strategies the child is using and write comments on the appropriate check list. The GOLD teachers created the report

card and the math and literacy recordkeeping (see pages 31 and 32).

As part of this record-keeping students keep their work in a variety of folders--such as math, handwriting, creative writing, science, center work. Students keep lists of what they've read during the year so I have a chance to see what they're doing and talk with them about it.

A key piece in our assessment is the Piagetian analysis of each child's cognitive skills (see seriation example and summary on pages 33 and 34). In the fall and spring, the town hires substitutes so that the teachers can test the children one-on-one. For me, this is important because the teacher has to be constantly aware of these thinking processes when working with kids. We use the Individual Piagetian Summary to record what we observe about a student's abilities in these areas at the beginning of the year. During the year, when a child reaches a new point, I note it and record it on the Piagetian Summary.

The record-keeping is the most time-consuming part but I believe in it because I know the children better and, in this program, it becomes an on-going assessment over three years. One of the biggest changes I've observed is a decrease in the level of student frustration and an improvement in self-esteem and student behavior. I think this happens because our process supports what kids can do, not what they cannot do. This program puts more responsibility on the teacher for curriculum and assessment. I have to ask, '*How do I know these kids know it?*'."

NOTE: These examples are *excerpts* from the actual forms

YEAR | TEACHER

STUDENT'S NAME

LITERACY SKILLS - GOLD PRIMARY AND INTERMEDIATE RECORD

Inter	Primary	Transitional
☐	☐	Reads independently using a wide range of books
☐	☐	Increases sight vocabulary
☐	☐	Continues to apply word strategies with unknown words, uses context to predict unknown words
☐	☐	Self-corrects while reading
☐	☐	Reads with comprehension characters, sequence, theme, motivation
☐	☐	Predicts, makes inferences related to story
☐	☐	Identifies favorite types of books, authors
☐	☐	Gathers information from non-fiction sources

		Writing
	☐	Name
	☐	Words
	☐	Phrases
☐	☐	Sentences
☐	☐	Stories
☐	☐	Formulates ideas
☐	☐	Composes complete thoughts
☐	☐	Edits and refines material

BEST COPY AVAILABLE

Student's Name _____ Year _____ Teacher _____

NOTE: These examples are *excerpted* from the actual forms

GOLD PRIMARY MATH STUDENT RECORD

1. NUMERATION FRACTIONS	1/2	1/4	1/3
Cuts whole/groups into parts; labels	☐	☐	☐
Recognizes symmetry	☐	☐	☐
Reads symbols	☐	☐	☐
Writes	☐	☐	☐

2. OPERATIONS	Concrete	Concrete or pictures with symbols	Symbols (+ - =)
SUBTRACTION			
Counts back mentally	☐	☐	☐
Separates sets	☐	☐	☐
differences from 10	☐	☐	☐
differences from 20	☐	☐	☐
Computes differences from 99	☐	☐	☐
without regrouping	☐	☐	☐
Computes differences from 99	☐	☐	☐
with regrouping	☐	☐	☐
Uses "0" in subtraction	☐	☐	☐
Knows minus is reverse of plus	☐	☐	☐

3. MONEY			
Recognizes coins	☐	☐	☐
Knows values of coins	☐	☐	☐
Counts coins	☐	☐	☐
Exchanges coins	☐	☐	☐
Makes change	☐	☐	☐

4. GEOMETRY	Concrete	Concrete or pictures with symbols	Symbols (+ - =)
Sorts, names plane shapes (2D),	☐	☐	☐
solids (3D)	☐	☐	☐
Recognizes shapes/solids in the environment	☐	☐	☐
Makes shapes/solids	☐	☐	☐
Reproduces patterns using shapes	☐	☐	☐
Makes and describes repeated patterns/tessellations	☐	☐	☐

printed with permission from
Joan Boykoff Baron, Connecticut
State Department of Education.

APPLIED MEASUREMENT IN EDUCATION, 4(4), 305-318
Copyright © 1991, Lawrence Erlbaum Associates, Inc.

Strategies for the Development of Effective Performance Exercises

Joan Boykoff Baron
Connecticut State Department of Education

This article contains four sections. The Nature of Assessment section explores the values manifested in the assessment, the clarity of the standards, the need to incorporate cognitive and motivational psychology, and the potential impact of the assessment on classroom instruction. The Properties of Effective Tasks section discusses the structural elements of effective tasks and issues related to students as problem solvers. The Making Tasks Meaningful and Engaging section calls for embedding tasks in "messy" real-world contexts. The Process of Developing Effective Performance Tasks section offers practical suggestions on the issues of who should be involved, sources of ideas for assessment, and how teachers are the critical change agents in achieving world class standards.

A fresh, cool breeze is beginning to be felt. Amidst the tremendous press for accountability being generated in all sectors of this country, there is some indication that classroom teachers may be asked to provide data about their students that will be used by policy makers to determine the success of our nation's schools. Why do I say this? At the June 1991 Annual Large Scale Assessment Conference, sponsored by the Education Commission of the States and The Colorado Department of Education held in Breckenridge, Colorado, there were several presentations about new state assessment programs in which teachers would be entrusted with the responsibility of being the primary source of important judgments about what their students know and can do. California talked about its new plan for using data from several embedded assessments as well as some on-demand testing (Carlson, 1991). Connecticut described one component of its Common Core of Learning Assessment program in which teachers would choose both the assessment tasks and the most appropriate time to administer them and,

Requests for reprints should be sent to Joan Boykoff Baron, Connecticut State Department of Education, Room 340, Box 2219, Hartford, CT 06145.

after scoring their students' work, would send it to the State Department of Education (Baron, 1991). Vermont outlined its plans to have its teachers not only collect their students' work in portfolios but also score it, using a moderation system adapted from one used in Great Britain (Brewer, 1991). In neighboring rooms at the Colorado meeting, representatives from the ARTS PROPEL program described how students' art, music, and creative writing have been scored successfully by Pittsburgh teachers (Camp, 1991), while researchers in the same school district updated us on some steps they are taking to validate and aggregate these judgments for policy use (LeMahieu, 1991).

As others have acknowledged, we have much to learn from our colleagues in different parts of the world (e.g., Great Britain and Australia) where there is a long tradition of entrusting teachers with providing valid judgments about students' work (see chapter VI in Raizen, Baron, Champagne, Haertel, Mullis, & Oakes, 1990). What is so promising about this new direction is that when teachers have access to rich assessment tasks and training in applying multidimensional scoring criteria, they feel equipped to make more nuanced and accurate judgments about their students' understandings and skills. Also, because assessments often serve to clarify educational goals, teachers feel more certain about what they are trying to accomplish. When beginning to implement complex performance exercises, many teachers find themselves asking the very empowering question, "What changes should I make in my curriculum and instructional strategies to enable my students to be successful on these performance exercises?"

This paper is written with both policy makers and teachers in mind. Therefore, it is written with one eye on accountability and the other on instructional guidance systems. It is written in the hope that one day soon data that is collected in the classroom in order to inform instructional decisions can also be used by policy makers to determine how well the educational system is succeeding. However, the realist in me acknowledges that this day may be several years away. Therefore, this paper is written for any state department of education personnel or classroom teacher who, for any reason, is interested in taking steps toward creating more effective performance assessment exercises. (Throughout this paper, I will use interchangeably the terms *performance assessment exercises* and *performance assessment tasks*.)

This article is divided into four sections. The first three sections contain a series of 19 questions designed to illuminate the characteristics of effective performance exercises. The first section, The Nature of Assessment, explores a broad set of issues surrounding the values manifested in the assessment, the clarity of the expectations and standards, the need to incorporate recent theory and research from cognitive and motivational psychology, and the potential impact of the assessment. The second section,

Properties of Effective Tasks, discusses both the structural elements of effective tasks and issues related to students as problem solvers. Effective tasks are seen as integrative, affording multiple solutions or solution paths, accessing students' prior knowledge, and, wherever possible, being loosely structured and allowing groups of students to work together. To foster problem solving, tasks should be sustained; they should allow for choice and control on the part of students, require students to both design and carry out investigations, and include opportunities for self-assessment and reflection. The third section, Making Tasks Meaningful and Engaging, stresses the importance of striving for as much authenticity as possible, for optimal levels of challenge, and for maximum levels of transfer of learning by embedding performance assessment tasks in "messy" real-world contexts. In the final section, The Process of Developing Effective Performance Tasks, some practical suggestions are offered related to the matters of who should be involved, how effective instructional tasks can be used as sources of ideas for assessment, and how teachers must be viewed as the critical change agents in achieving world class standards.

THE NATURE OF ASSESSMENT

Assessment can legitimately be viewed as the manifestation of a system's educational values. Therefore, in designing a set of assessment tasks, there are six important questions that should be revisited frequently. These will be discussed in turn.

1. "Does my set of assessment tasks embody what I value both as a representative of my content discipline(s) and the educational community that I serve? That is, if a group of curriculum experts in my field and a group of educated citizens in my community were to use my assessment tasks as an indicator of my educational values, would I be pleased with their conclusions? And would they?"

Suppose we were setting out to design a new educational assessment program and we started by talking with people about their views of today's schools. What would we hear? From parents, we might hear, "Stop dumbing down the curriculum. Our schools don't expect enough from our children". Curriculum leaders would tell us that "Less is more"—students should understand a small number of big ideas in each discipline. From business and industry would come the plea: "We want students to make informed decisions and be able to use their knowledge and skills to solve real-world problems." From everyone in unison we might hear, "We want students to be self-confident learners who can read, write, and do arithmetic

as well as think, make decisions, communicate effectively, and work well together. Both our educational programs and our assessment systems should respond to these concerns.

2. "When students prepare for my assessment tasks and I structure my curriculum and pedagogy to enable them to be successful on these tasks, do I feel assured that we are all engaging in authentic and ecologically valid activities?"

If teaching to the test or studying for the test results in actions which make students more effective as test takers only, rather than as genuine or authentic readers, mathematicians, writers, historians, artists, problem solvers, etc. then the tasks are falling short and need revision (Frederiksen & Collins, 1989). We must keep in mind that we are ultimately preparing our students to function as effective citizens in today's and tomorrow's world. Therefore, the greater the synchrony between what we value and what we assess, the more effective will be the resultant educational system.

3. "Do my assessment tasks reflect an understanding of human psychology with its recent advances in cognition, learning theory, motivation, and instruction?"

Three related concomitants of effective tasks are that they foster students' becoming active, engaged, and responsible for their own learning. By using psychological theory and research, tasks can be designed to increase the likelihood that this can happen (Baron, 1990; Baron, Forglone, Rindone, Kruglanski, & Davey, 1989; Ronning, Glover, Conoley, & Witt, 1987). Several specific questions in later sections of this article will incorporate knowledge gained from the field of psychology.

4. "What content, processes, and dispositions should my tasks assess?"

Content is primary. We should not set out to design "content-free" assessment tasks. An effective performance exercise must incorporate the big ideas and essential concepts, principles and processes in a discipline. Therefore, in attempting to answer this question, task developers will be forced to look at the forest as well as the trees. Most disciplines have already identified small set of big ideas and processes that are critical to developing deep understanding of a given content area. These are available from several national associations (e.g., the American Association for the Advancement of Science, 1989; the National Council of Teachers of Mathematics, 1988). In addition, some state curriculum and assessment frameworks have begun to help teachers map their content onto appropriate big ideas and processes (see recent California curriculum frameworks and Kentucky assessment frameworks). However, by way of warning, these documents provide several good examples, but they do not provide a comprehensive set of performance exercises. The job of creating a bank of tasks to assess the big ideas still remains to be done.

5. "Do my tasks clearly communicate my standards and expectations to my students?"

One of the advantages of using performance exercises is that they communicate clearly to students what is expected of them, thereby allowing students to internalize the criteria for successful performance. Optimally, as they prepare for assessment tasks, students should be fully aware of the criteria for successful performance. This can be accomplished by sharing openly with them, throughout the school year, the scoring criteria that will be used to assess the quality of their work. In this way, students can self-monitor the quality of their own work as well as receive corrective feedback from the teacher. In short, when these activities occur, the multifaceted dimensions of effective performance become an integral part of the conversations in the classroom. (If the students have not had continual practice with using the criteria throughout the year, at the very least and as a last resort, the criteria should be shared simultaneously with the performance exercise.)

6. "In attempting to learn what my students know and can do, am I making the best use of performance assessment?"

Performance assessments take more time to administer and score than multiple-choice tests do. Therefore, performance assessments should be used to assess those understandings and processes that cannot be tapped adequately by multiple-choice tests. If our agenda is to produce students who have a deep understanding of important content, a sense of personal efficacy, and the abilities to think critically, make decisions, solve problems, communicate, and collaborate with others, then our assessments should reflect those values. To achieve these ends, performance assessments are far better suited than multiple-choice tests.

PROPERTIES OF EFFECTIVE TASKS

This section is divided into two parts—the Structure of Performance Tasks and Students as Problem Solvers. Structurally, effective tasks are integrative; they afford multiple solutions or solution paths, access students' prior knowledge and, wherever possible, are loosely structured and allow groups of students to work together. To foster problem solving, tasks should: be sustained, allow for choice and control on the part of students, require students to design and carry out investigations, and include opportunities for self-assessment and reflection.

The Structure of Effective Problems

In this section, tasks are envisioned as opportunities for students to make connections among disparate pieces of the curriculum. They are seen as providing opportunities for individual students or groups of students to employ different perspectives and several possible solutions.

7. "Are some of my tasks rich and integrative, requiring students to make connections and forge relationships among various aspects of the curriculum?"

Consistent with the focus on big ideas and essential understandings described earlier, the purpose of effective tasks is to allow students to foster and display a depth of understanding rather than a breadth of understanding, permitting them to make connections among previously fragmented knowledge and skills. Effective tasks also allow students to be able to tell a whole story. Today, so much of what students learn is highly fragmented and disjointed. Because they lack rich conceptual networks to organize their knowledge, they quickly forget what they learn (Bransford & Stein, 1984; Bruner, 1990). When students are encouraged to think about a whole problem, as they would be when designing and carrying out a complete investigation, their knowledge is more unified and more likely to be retained over time.

8. "Have I included some messy, loosely structured problems in which students have to first structure the problems before beginning to solve them?"

Problem finding and problem formulation are skills that are highly desirable in the workplace as well as in higher education. Therefore, it may be optimal to include also some tasks that require students to formulate their own problems (Greeno, 1978; Resnick, 1989; Schoenfeld, 1976).

9. "Do my tasks have either multiple solutions or solution paths, and do they encourage diverse perspectives?"

One big advantage of performance tasks over multiple-choice questions is that they permit many different approaches. Therefore, one makes best use of the performance assessment when the problems chosen can be solved using a variety of strategies and/or if they allow several justifiable solutions. Some tasks may explicitly require students to take on multiple perspectives in solving the problem. Still others might even call for deep sense critical thinking, in which students engage in dialogical thinking. Such tasks require students to take a particular point of view, challenge it from a different perspective, and then respond to the challenges from the first point of view (see Paul, 1987). This feature of task structure is particularly well suited to requiring students to engage in pluralistic thinking in which a problem can be viewed from the perspectives of several cultural contexts.

10. "Are my tasks structured to encourage students to access their prior knowledge and skills when solving problems?"

This has at least three advantages. First, students can use their existing knowledge and skills to "prime the pump" and help them to enter a novel situation. Second, thinking about what they already know about a subject may help them integrate their current findings into a more tightly structured conceptual network. Finally, when new findings are incompatible with their

prior beliefs and concepts, the discovery may provide a first step for reconciling these inconsistencies (see Borkowski, Carr, Rellinger, & Pressley, 1990; Brown, Bransford, Ferrara, & Campione, 1983; Paris & Winograd, 1990; Pearson & Raphael, 1990).

11. "Do some tasks require students to work together in small groups to solve complex problems?"

Schools should provide opportunities for students to work together to solve complex problems. However, the reader is cautioned that not all individual tasks are equally successful as small group tasks. In order to warrant the use of several hours of group time, the task should allow for a diversity of approaches to be considered, and either be sufficiently large so that several students are needed to divide the work or else it should allow opportunities for members of the group who begin with partial understandings to deepen their understandings through group conversation (Aronson, Blaney, Stephan, Sikes, & Snapp, 1978; Hibbard & Baron, 1990; Vygotsky, 1978/1935). If tasks are loosely structured and complex, groups will often benefit from the presence of multiple perspectives. The diversity of viewpoints is not mere window dressing; rather, it generally maximizes the group's likelihood of success.

Students as Problem Solvers

The need to produce effective problem solvers is being championed in every segment of our society. In this section, task developers are urged to think about providing sustained problems affording both choice and reflection. The design and carrying out of investigations are seen as viable approaches to enhancing problem solving.

12. "Do some problems require sustained work?"

Some tasks should require several days of work during which students visit and revisit the problem in several iterations. For some problems, it would be appropriate to allow students to work intermittently over several weeks or even months. Examples may be found in science, in which students study an ecological niche over many seasons, or in the arts, where students create several renditions of a subject in different media or from different historical or cultural perspectives (Quellmalz, 1987; Wolf, 1989).

13. "Do some tasks allow students a degree of choice and control over the course of action needed to solve problems and conduct investigations?"

In fact, are some problems such that they can be formulated by the students? This idea is designed to give the students the maximum possible responsibility for planning and decision making as well as a sense of efficacy and personal control (Bandura, 1982). When students leave school, they will not have a teacher close by to help them solve the dozens of

BEST COPY AVAILABLE

problems they will face on any given day. School is the place where students should be encouraged to make their own choices in complex problem-solving situations. If we want to produce a creative and innovative work force and citizenry, we will want to give students practice in problem finding as well as problem solving (Charles & Silver 1988; Nation Council of Teachers of Mathematics, 1988).

14. "Do some tasks require students to design and carry out their own investigations?"

Although students are often asked to carry out investigations in science classrooms, they rarely have a chance to design their own investigations. The design element is vital for the skills we value in society. The following five questions highlight desirable aspects of investigations:

> Do some tasks require students to make estimates and predictions before conducting their investigations?
>
> Do some tasks require students to collect, analyze, and portray their data?
>
> Do some tasks require students to explain their findings and cite evidence for their conclusions?
>
> Do some tasks require students to identify their assumptions and possible sources of error?
>
> Do all of my tasks require students to communicate their findings in writing and orally?

Students have shockingly few opportunities to describe orally what they know and can do. Rich problems which give rise to many different solutions also provide interesting opportunities for students to share their approaches with one another. Informal oral presentations are very useful for this purpose and should become a routine part of performance assessment. Tasks should also encourage the display of numerical data in a variety of formats (e.g., charts, tables, and graphs).

15. "Do some of my tasks require self-assessment and reflection on the part of students?"

Students should not only be problem solvers, but they should be encouraged to reflect upon themselves as problem solvers. This reflection can be built into the performance exercise by asking students to assess their own work (Wolf, 1989). Initially, this process is facilitated by making available to the students the scoring criteria along with discussions of the meaning of the criteria. Since the ultimate goal is for the students to

internalize the criteria eventually, teachers can experiment with the best balance of providing and withholding criteria. For example, on some occasions, teachers may wish to remove the criteria after substantial practice with them to see whether students are disposed to use them on their own.

MAKING TASKS MEANINGFUL AND ENGAGING

Unfortunately, so much of schooling has so little meaning for the large numbers of students who remain unengaged. Even the great majority of students who come to school and do enough work to be promoted and eventually graduate do not emerge from high school as the active problem solvers we need. Therefore, this section has two parts: Making Tasks Meaningful for Students and Situating Tasks in Challenging Real-World Contexts.

Making Tasks Meaningful for Students

In making tasks more meaningful for students, authenticity and degree of challenge are two important ingredients.

16. "Are my tasks likely to have personal meaning for the students?"

Some have referred to this as authenticity, claiming that students should be asked to solve problems that are real for them, for example, problems about whose solutions they care. The real-world problems described below are advocated because they are likely to have more meaning for students.

17. "Are my tasks sufficiently challenging for the students?"

There should be a careful blending of the familiar and the novel so that students feel challenged and yet efficacious when facing the task. Moderate degrees of challenge are considered best at the beginning. The degree of challenge can be increased as students are better able to cope with some frustration and delay of gratification.

Situating Tasks in Challenging Real-World Contexts

There are several reasons why it is desirable to situate problems in real-world contexts. Two of the more important ones are that students are more likely to find such problems engaging, and they are more likely to facilitate transfer by demonstrating to students that their knowledge and skills are useful in solving real problems (Brown, J. S., Collins, & Duguid, 1989; Rogoff & Lave, 1984).

18. "Do some of my tasks provide problems that are situated in real-world contexts and are appropriate for the age-group solving them?"

We want students to be able to solve real-world problems. And we want them to do it thoughtfully, using their prior knowledge and skills. Therefore, wherever possible, assessment tasks should make use of situations that are relevant for the students who will solve them. (However, real-world contexts should not be considered essential for all problems. Several teachers report that their students can become extremely absorbed in solving some very abstract problems when the content is challenging.)

19. "Do some tasks allow for transferring the understandings gained and generalizations made in the present task to other related tasks?"

In this way students become more adept at appropriately using their declarative and procedural knowledge (i.e., knowledge and skills) in a variety of situations. One of the problems noted by teachers and psychologists is that students may know a lot and have a large repertoire of skills, but they are insecure about knowing when to use them (Cormier & Hagman, 1987). If students have practice in solving loosely structured real world problems and are asked explicitly to transfer their understandings to other contexts within the structure of the task, we may make some strides in helping students to transfer their knowledge and understandings naturally in other situations.

THE PROCESS OF DEVELOPING EFFECTIVE PERFORMANCE TASKS

This section is concerned with how one might go about finding and/or developing effective performance tasks.

Filling the Need for Effective Performance Tasks

Unfortunately, there do not yet exist task banks of performance exercises that teachers, state department of education personnel, or researchers thinking about national examining systems can peruse to find the appropriate assessment tasks for a particular occasion. Although there is increasing interest in sharing tasks across school districts and between states, there are disappointingly few tasks currently available to share. We are probably still several years away from being able to do so. That means that everyone reading this volume can play an important role. It is almost dizzying to speculate about how much further advanced we could be if each reader of this journal were to develop just one performance assessment

exercise this year. We would then be faced with the enviable challenge of developing appropriate mechanisms for sharing them with like-minded colleagues.

Getting Started: Who Should Be Involved?

If we are committed to developing tasks which reflect deep understanding of content, critical thinking, problem solving, communication, and collaboration, teachers working together with curriculum experts make the best starting team. This team will include curriculum experts who can identify big ideas within each discipline and classroom teachers who can identify authentic and engaging tasks likely to be motivating for their students. After some tasks and preliminary scoring criteria have been roughed out, it may be useful to have them reviewed and critiqued by some assessment specialists, cooperative learning experts, and psychologists informed in cognitive, motivational, and learning principles. Students can be involved in the development process by providing for analysis both their responses to the exercises and their reactions to various aspects of the task (e.g., clarity, level of challenge, degree of engagement, etc.).

Looking at Instructional Tasks for Some Good Ideas

Several teachers who have been working on the development of performance exercises have noted some striking similarities between effective assessment tasks and effective curriculum tasks. After a decade of developing assessment tasks in more than a dozen areas, my colleagues and I would agree that they can be very similar (Baron, et al., 1989). Good instructional tasks embody virtually all of the criteria delineated in this article. There are, however, three important differences between effective instructional and assessment tasks. The first is that assessment tasks must include a set of scoring criteria, whereas (unfortunately) instructional tasks often do not include one. The second is that assessment tasks should be highly integrative in nature. They should be viewed as culminations of a large number of instructional tasks that preceded them. Assessment tasks are intended as occasions for connections to be made and bridges of understanding to be built. This is not to suggest that similar connections should not be encouraged in instructional tasks. However, instructional tasks are often designed to develop new skills or learnings rather than to synthesize and integrate prior learnings. The third difference is in the role of the teacher. Whereas the teacher's role in the instructional task is to mediate the student's learning through selectively administered cues (i.e., the role of catalyst or coach), the teacher's role in the assessment task is viewed as

virtually "hands off." Students should be allowed to work alone or in a small group to solve problems without the teacher's help. On balance, because they are more similar than different, good instructional tasks may offer some interesting possibilities for designing effective assessment tasks.

Others have made the point that assessment is an occasion for learning (Wiggins, 1989; Wolf, Bixby, Glenn, & Gardner, 1991). This is because when students are given a challenging problem they are able to see relationships and connections that were not obvious to them before. The learning that takes place is a function of the interaction between the student(s) and the task or the students and each other. It does not take place because the teacher is actively mediating the students' thinking.

THE TEACHER AS THE CRITICAL CHANGE AGENT: FORGING COALITIONS

In the end, it matters little what policy makers say or do or what national curriculum groups espouse unless teachers find their ideas alluring enough to implement. In striving for world class standards, the focus for change as well as the locus for change must be the classroom. The time is long overdue for coalitions of teachers, curriculum developers, assessment specialists, and educational psychologists to create appropriate contexts in which to assess and foster the abilities of our students to deeply understand important content and to think, problem solve, communicate, and collaborate. The time is ripe to work together to create for our teachers and our students rich assessment opportunities in which students can "put it all together" and self-monitor their own learning. Teachers can use these data to determine what students know and can do, and fashion appropriate learning environments and instructional episodes. And before too long, we will figure out creative ways to aggregate these rich data for policy use at the state and national levels.

ACKNOWLEDGMENTS

Many of the ideas in this article resulted from my work on the Connecticut Common Core of Learning Assessment Program funded by the Connecticut State Department of Education and the National Science Foundation (SPA-8954692).

The thoughts expressed in this article are my own and do not necessarily represent the views of the funding agencies.

I thank Reuben M. Baron for his helpful comments on an earlier draft.

REFERENCES

American Association for the Advancement of Science. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology.* Washington, DC: Author.

Aronson, E., Blaney, M., Stephan, C., Sikes, J., & Snapp, M. (1978). *The jigsaw classroom.* Beverly Hills, CA: Sage.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37,* 122-147.

Baron, J. B. (1990). Performance assessment: Blurring the edges among assessment, curriculum, and instruction. In A. B. Champagne, B. E. Lovitts, & B. J. Calinger (Eds.), *This year in school science: Assessment in the service of instruction* (pp. 127-148). Washington, DC: American Association for the Advancement of Science.

Baron, J. B. (1991, June). *Using performance measures to assess the Connecticut Common Core of Learning.* Paper presented Education Commission of the States/Colorado Department of Education Conference on Alternative Assessment, Breckenridge, CO

Baron, J. B., Forgione, P. D., Rindone, D. A., Kruglanski, H., & Davey, B. (1989, April). *Toward a new generation of student outcome measures: Connecticut's Common Core of Learning Assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Borkowski, J. G., Carr, M., Rellinger, E., & Pressley, M. (1990). Self-regulated cognition: Interdependence of metacognition, attributions, and self-esteem. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (Vol. 1, pp. 53-92). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bransford, J. D., & Stein, B. S. (1984). *The ideal problem solver: A guide for improving thinking, learning and creativity.* New York: W. H. Freeman.

Brewer, R. (1991, June). *Portfolio assessment – Findings from research and practice.* Paper presented at the Education Commission of the States/Colorado Department of Education Conference on Alternative Assessment, Breckenridge, CO.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), *Carmichael's manual of child psychology* (Vol. 1, pp. 77-166). New York: Wiley.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning *Educational Researcher, 18*(1), 32-42.

Bruner, J. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Camp, R. (1991, June). *Portfolio assessments – Findings from research and practice.* Paper presented at the Education Commission of the States/Colorado Department of Education Conference on Alternative Assessment, Breckenridge, CO.

Carlson, D. (1991, June). *Performance assessment in California.* Paper presented at the Education Commission of the States/Colorado Department of Education Conference on Alternative Assessment, Breckenridge, CO.

Charles, R. I., & Silver, E. A. (Eds.). (1988). *The teaching and assessing of mathematical problem solving* (Vol. 3). Reston, VA: Lawrence Erlbaum Associates and the National Council of Teachers of Mathematics.

Cormier, S., & Hagman, J. (Eds.). (1987). *Transfer of learning: Contemporary research and applications.* Orlando, FL: Academic.

Frederiksen, J. R., & Collins, A. L. (1989). A systems approach to educational testing *Educational Researcher, 18*(9), 27-32.

Greeno, J. (1978). A study of problem solving. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 1, pp. 13-75). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hibbard, K. M., & Baron, J. B. (1990, April). *Assessing students working in groups.* Paper

from cooperative and collaborative learning. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

LeMahieu, P. (1991, June). *Reporting authentic assessment: Is calamity brewing?* Paper presented at the Education Commission of the States/Colorado Department of Education Conference on Alternative Assessment, Breckenridge, CO.

National Council of Teachers of Mathematics. (1988). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Paris, S. G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (Vol. 1, pp. 15-51). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Paul, R. W. (1987). Dialogical thinking: Critical though essential to the acquisition of rational knowledge and passions. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 127-148). New York: W. H. Freeman.

Pearson, P. D., & Raphael, T. E. (1990). Reading comprehension as a dimension of thinking. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (Vol. 1, pp. 209-240). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Quellmalz, E. S. (1987). Developing reasoning skills. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 86-105). New York: W. H. Freeman.

Raizen, S. A., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I. V. S., & Oakes, J. (1990). *Assessment in science education: The middle years* (pp. 65-75). Andover, MA: The Network.

Resnick, L. B. (1989). Teaching mathematics as an ill-structured discipline. In R. I. Charles & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (Vol. 3, pp. 32-60). Reston, VA: Lawrence Erlbaum Associates, Inc. and the National Council of Teachers of Mathematics.

Rogoff, B., & Lave, J. (1984). (Eds.). *Every day cognition: Its development in social context*. Cambridge, MA: Harvard University Press.

Ronning, R. R., Glover, J. A., Conoley, J. C., & Witt, J. C. (1987). (Eds.). *The influence of cognitive psychology on testing*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Schoenfeld, A. (Ed.). (1976). *Cognitive science and mathematics education*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Trans. and Eds.). Cambridge, MA: Harvard University Press. (Original work published 1935)

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wolf, D. (1989). Portfolio assessment: Sampling student work. *Educational Leadership, 46*(7), 35-39.

Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Research in Education* (Vol. 17, pp. 31-74). Washington: American Educational Research Association.

BEST COP

Y AVAILABLE

SECTION II: Science Assessment K-6

Reprinted with permission from NSTA Publications, copyright November/December 1992, Science and Children.

Evaluating Elementary Science

By Rodney L. Doran, Douglas Reynolds, Janice Camplin, and Nicholas Hejaily

Since the advent of the activity-oriented science programs of the 1960s, educators have been concerned about the need for appropriate assessment instruments. Today, New York State has addressed that need with an innovative system that includes authentic science assessment.

Make a Change

The changes in New York didn't happen overnight. For decades (since 1958), elementary science instruction had followed guidelines specifying topics for each grade level. There was no state assistance with local program development, nor were there any state elementary science tests.

Then, in 1985, four important changes took place. First, a new state syllabus established outcomes for science content, skills, and attitude. Second, teacher supplements to that syllabus featured sample program materials. Third, an assessment device, the Elementary Science Program Evaluation Test (ESPET) was proposed to assess the effectiveness of the science instructional program in each school. Fourth, the state developed a statewide elementary science mentor network of experienced teachers and administrators trained to support the development of local school programs and to administer and score ESPET. Without all four components of the system (syllabus, program, assessment, and mentor network), it is unlikely that positive, sustained change in elementary science could have occurred.

In May 1989, 211,000 fourth-grade students, handicapped and nonhandicapped, enrolled in public and nonpublic schools in New York, took the first ESPET. Based on the outcomes of Levels I and II of the new syllabus (grades K–4), the untimed, 45-item multiple-choice objective test measured science content and inquiry skills. In addition, many school districts administered one or more of the optional survey components that included a 20-item science attitudes survey, a 25-item program environment survey, a 36-item teacher survey, a 32-item administrator survey, and a 12-item parent/guardian survey.

Testing Manipulative Skills

New York State was the first in the nation to administer a large scale manipulative skills test statewide to all students' at a given grade level. The test consisted of tasks at a series of five locations (stations) set up in a room by the school's elementary science mentor (ESM).

Station setups, equipment lists, and verbal directions were standardized across the state according to an "administrator's guidebook" that each mentor had received in a workshop. These guidelines enabled the test administrators to check materials quickly, just before students moved to the next station, and helped make the test valid across the state. Figure 1 shows the materials list, station setup diagram, and directions for preparing "Station Two—Water on Objects."

Figure 2 shows a classroom setup for 25 students. Students were able to move from station to station with minimal instructions. Because they were kept busy at their own stations, the children had little temptation to observe others' work. Also, the stations were staggered, so any "observing" would be of some other task.

Trial Testing

We began by collecting existing assessment ideas from as many sources as possible. To develop the skills test of ESPET, we examined the following sources:

- the Second International Science Study (SISS),
- the Assessment of Performance Unit (APU),
- the National Assessment of Educa-

tional Progress (NAEP),

- and commercial material and curriculum projects (SCIS and SAPA, for example).

Doran's (1990) review of "practical" tests from various national and international surveys and Hein's (1990) collection of papers on the assessment of hands-on programs also came into play.

The SISS had developed and administered skills tests for fifth graders, so we modeled our initial tasks on their material and modified tasks from other sources accordingly. Our model asked a series of questions connected to a set of materials or a common problem; involved a set of safe, easily accessible materials and equipment; and included a set of scoring guidelines with point values allocated to each question.

Producing a set of tasks for statewide testing purposes that will work well in a wide variety of classrooms takes at least two years to develop and refine. Because these tests are assessing the outcome of a specific curriculum (school, district, or state), the goals and objectives for that curriculum must exist in a published form and be considered at each stage of development.

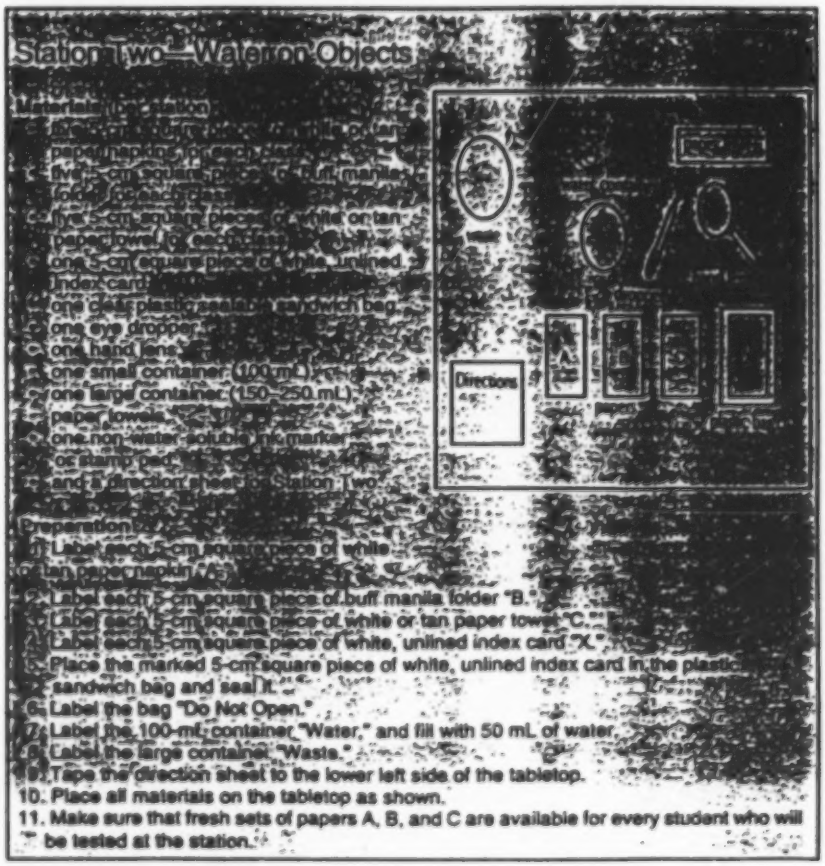
A team consisting of the regional elementary science mentors from western New York, specialists from the science and testing bureaus of the state's Education Department, and a university science educator was charged with shepherding these tasks from inception to statewide implementation. The creativity and wisdom from their teaching and mentoring experiences were essential at every stage of development.

As each new task was proposed, one mentor was assigned to refine it and modify it to a common format. If it seemed workable, the mentor tried the task with a few students, checking the materials' suitability and the time needed to complete the task. After making necessary revisions, the team tried a sample of tasks with 10–15 students.

At that point, the pool of tasks were analyzed according to skills assessed, grade level of language used, material accessibility and "mess quotient." After screening and subsequent revisions, these tasks were combined in a mock-up with the same number and length of time as desired for the statewide test. The mock-up was tried in six to eight schools (one class each) with verbal directions, test booklets, timing, and so on. The mentor (or a classroom teacher) administered the test. Other experienced mentors and representatives of the state Education Department observed the session, looking for difficulties with reading, equipment, and time constraints.

We tested about twice as many tasks as we would need for the ESPET. Based on student responses and on observations by teachers and test administrators, we assembled recommendations for modification. The items that assessed a broad range of skills and had the best testing characteristics were then assembled for statewide field testing.

Figure 1. Station Two's materials list, setup diagram, and directions.


Station Two—Water on Objects
Directions
Preparation
1. Label each 5-cm square piece of white or tan paper napkin "A."
2. Label each 5-cm square piece of buff manila folder "B."
3. Label each 5-cm square piece of white or tan paper towel "C."
4. Label each 5-cm square piece of white, unlined index card "X."
5. Place the marked 5-cm square piece of white, unlined index card in the plastic sandwich bag and seal it.
6. Label the bag "Do Not Open."
7. Label the 100-mL container "Water," and fill with 50 mL of water.
8. Label the large container "Waste."
9. Tape the direction sheet to the lower left side of the tabletop.
10. Place all materials on the tabletop as shown.
11. Make sure that fresh sets of papers A, B, and C are available for every student who will be tested at the station.

The Education Department's Bureau of Testing selected 20 schools (a sample of 1,053 fourth graders) to represent the diversity of schools within the state. These schools were invited to participate in the next step—the statewide field test. This field testing was done exactly as it would be in subsequent statewide implementation, in terms of directions to be read by the test administrator, questions asked, student instructions, and materials and equipment kits. Following the field testing, we made final adjustments to the instrument and also established norms for interpreting the results.

Scoring Procedures

Because teams of teachers at each building or district would be scoring the tests, we developed a clear and easy-to-use set of scoring procedures:

- a small number of points were allocated to each item;
- a scoring guide clearly stated the criteria for assigning points;
- and examples of answers were given for each point value.

A set of scoring procedures was drafted, revised, and trial-tested prior to use in the statewide field test. Each station had several questions for the student to answer. Each question had a maximum point value between one and three. As the maximum number of points possible for any question was three, the deliberation by the raters was minimal. The point value for each station was four or five, with a maximum score on the skills test of 22 points. The rating guide provided as much help as possible in listing sample correct answers, but scorers were instructed to give appropriate credit for answers conveying the same general meaning as those listed. Such

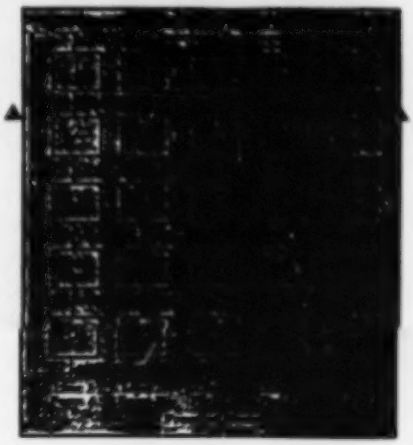


Figure 2. Sample classroom setup.

a provision is critical if scorers are to avoid the "single answer mentality" that pervades assessment.

The rating/scoring guide is organized into three parts—an overall set of procedures, criteria and acceptable answers for each item, and sample student answers illustrating each point value for each item.

The criteria and acceptable answers for the first question at Station Two are shown below, along with a student example from the rating guide.

What happened to the drop of water on each piece of paper?

Criterion: The student correctly describes what happened to the drop of water on each piece of paper.
Maximum score: three points
Samples of acceptable answers:
On papers A and C, the drop of water
  was absorbed
  soaked in
  spread out
  got bigger
  expands
  fills up/makes squares or blocks
  goes through
On paper B, the drop of water
  was not absorbed
  sits on top
  stays in ball
  stays the same
  doesn't spread
  stays a drop
  won't go through
  bubbles on top
Number of Credits
What happened to the drop of water on each piece of paper?
+1 On paper A, the water drop soaked in.
+0 On paper B, the drop of water probably soaked in and spread out.
+1 On paper C, the drop of water soaked in.
Comment: The student correctly described what happened to the drop of water on papers A and C, but not on paper B.

Did Scorers Agree?

Seven teachers who had been trainers for the scoring workshops traveled to Albany to re-rate 3,949 of the May 1989 ESPET manipulative skills tests. The number of discrepancies was tallied by school, by station, and by item within each station.

Overall, the two ratings agreed more than 90 percent of the time. The level of agreement, however, did vary considerably by task and by question within task. For example, an item on measuring mass involved only clear, simple scoring criteria and led to few discrepancies; while a request that students "write a statement about electricity" called upon scorers' subjective judgement and resulted in a high number of discrepancies.

A compilation of the scoring difficulties and recommendations for remediation was sent to schools whose papers were re-rated. Schools with more than 10 percent total discrepancies were recommended for a scoring workshop.

The Results Are In

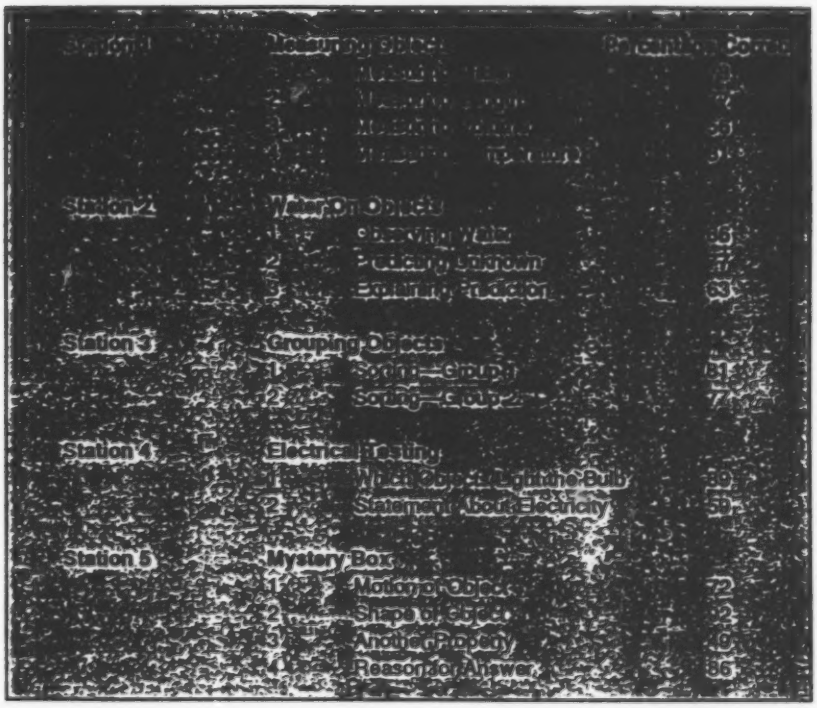
Schools were required to send to the state Education Department a summary of their ESPET performance on a form called the "Comprehensive Assessment Report" (CAR). Similar school summary data were collected on CAR for other state-administered tests. These data were then compiled so schools could see how their perfor-

(continued on page 63)

(Evaluating, continued from page 35)

mance compared with other schools in their county, with similar-sized schools, and with all the state's schools.

Performance on the manipulative skills test at each school was also viewed within the context of the statewide data. The data shown below reflect the percentage of students who answered correctly for each question of the skills test (May 1989 data).



Looking at the data by station and by item, we can make some generalizations. At the first station, students scored relatively well on simple measuring tasks, with 70–80 percent correctly determining mass, length, and temperature with a centimeter scale, thermometer, and pennies (nonstandard mass units).

Measuring volume proved a bit more difficult; students had to pour water into an unmarked glass or measuring cup and then compare with either a line on the glass or the scale on the measuring cup. This series of steps made volume measurement more complex than the other measurements. Some students incorrectly used the centimeter scale or the thermometer for measuring volume.

At the second station, students had to record their observations of water drops placed on three kinds of paper. They were very successful at this task; 86 percent earned full credit. The second item asked students to predict what would happen to a water drop on a piece of paper within a sealed clear-plastic bag. The "unknown" paper was similar to one of the "tested" papers, and students were relatively successful with this prediction (77 percent correct). The third item probed the reasoning behind their prediction; fewer children were able to do this than predicted correctly (63 percent).

At the third station, students were asked to sort objects (seeds) into two groups based on a common property and then state the shared characteristic for each group. Approximately 80 percent of the students sorted the seeds into two groups and described the shared characteristic.

At station four, most students benefitted from previous work with batteries, bulbs, and wires. They were skilled in checking the five objects (wire, paper clip, spoon, foil, and toothpick) in a circuit to see which one made the bulb light. The scoring for this item was as follows:

- 3 points—5 objects tested correctly
- 2 points—4 objects tested correctly
- 1 point—3 objects tested correctly
- no score—less than 3 objects tested correctly

Most students were successful in testing these objects. The second item—writing a statement about electricity and all the objects tested—was considerably more difficult. The average score for this item was 59 percent. Many students wrote only about the objects that made the bulb light rather than including a statement about *all* the objects.

The fifth station required students to infer properties of objects within a sealed box from the senses of hearing and touch. The students were more successful inferring the shapes of the objects (92 percent) than their motion (72 percent). "Shape" is a widely used property, more familiar than motion. Another item at the station asked, "What is another property of an object in the box?" This also proved to be difficult (49 percent correct). Finally, students were asked to explain how they estimated the number of objects in the box. As most students "heard" collisions between objects, they were very successful at this task (86 percent).

Follow a Leader

It *is* possible to assess manipulative skills associated with an elementary science program, but it does not happen in a vacuum. In New York State,

we built upon a revised syllabus, mandates for local program development and state level assessment, and an elaborate network of teachers training teachers (mentors). We hope individual teachers, school districts, and states will continue building upon this foundation.

Resources

Doran, R.L., and Meng, E. (1990). What research says about appropriate methods of assessment. *Science and Children, 28*(1), 42–45.

Hein, G. (1987). The right test for hands-on learning? *Science and Children, 25*(2), 8-12.

____ . (Ed.). (1990). *The assessment of hands-on elementary science programs.* Grand Forks, ND: North Dakota Study Group.

Kanis, I., Doran, R., and Jacobson, W. (1990). *Assessing science laboratory skills at the elementary and middle/junior high school levels.* New York: Second International Science Study.

New York State Education Department. (1985). *Elementary science syllabus.* Albany, NY: Author.

____ . (1988a). *Guide to program evaluation K–4.* Albany, NY: Author.

____ . (1988b). *Rating guide for the manipulative test: Grade 4, Form X.* Albany, NY: Author.

____ . (1990). *Program evaluation test in science—Direction for administering and scoring.* Albany, NY: Author.

RODNEY L. DORAN is professor of science education at the State University of New York at Buffalo. DOUGLAS S. REYNOLDS is the chief of the Bureau of Science Education with the New York State Education Department in Albany. JANICE CAMPLIN is a former curriculum coordinator at Lake Shore (New York) Schools. NICHOLAS HEJAILY is director of science for the Williamsville (New York) Central Schools.

84

Reprinted with permission from EDUCATIONAL LEADERSHIP, February 1993.

Science for All: Getting It Right For the 21st Century

Kenneth M. Hoffman and Elizabeth K. Stage

National standards in curriculum, teaching, and assessment—to be published in the fall of 1994—will translate the vision of "science for all" into concrete direction for achieving it.

In December 1892, 18 men met at the University of Chicago to advise the "Committee of Ten" on science preparation needed for college admission. The consultants were teachers from high schools and prep schools and faculty at public and private colleges and universities. Their consensus was that at least one year of biology, followed by one year of chemistry and one year of quantitative physics, would best prepare young people to grow up to be just like them (National Education Association 1894).

Recently, in December 1992, 600 men and women—mathematics and science teachers, supervisors, state coalition and systemic initiative directors, assessment reformers, governors' education aides, and others—gathered in Washington, D.C., to discuss preliminary working documents of the National Committee on Science Education Standards and Assessment (NCSESA). These materials address "Science for All," a challenge that our nation is finally, 100 years later, ready to embrace.

The Legacy of the Committee of Ten

There's a lot to like about the 1892 reports to the 10 college and university presidents, including a recommendation "that the laboratory record should form part of the test for admission to college." What's not easy to like is that the content recommendations set the high school curriculum that remains in place today for nearly all students. This has led to the current situation: *some* science for *some* students.

As the body of scientific knowledge has exploded, high school courses have become cluttered with so much new vocabulary—often exceeding that of foreign language courses—that terms can only be memorized rather than understood. To prepare students for this onslaught of disconnected

facts, junior high courses have often imitated high school courses, with levels of abstraction and quantification that go beyond the intellectual capacity of young people. Junior high students, too, have learned to succeed by memorization. Since memorized work is easily forgotten, teachers at each level teach as if students' minds are empty. Thus, the expectations for elementary school children are usually minimal: "Keep their curiosity alive."

The elementary curriculum depends largely on the interests of teachers, only a quarter of whom feel "well qualified" to teach science (Weiss 1989). Is it any surprise, then, that although 70 percent of elementary students say they are interested in science (Weiss 1989), by the time they reach high school, science enrollments drop by more than one half each year? Only 20 percent of high school students nationally take the final course in physics recommended in 1892 (Blank and Dalkilic 1990).

Challenges to the Status Quo

What's wrong with that? Since only 3 or 4 percent of the work force is engaged in science and engineering (U.S. Department of Labor 1992), why do all of our citizens need to learn science? Concerns about competitiveness in the global economy are fueling the renewal of science and mathematics education. The business community demands entry-level workers who are able to think and solve problems. Regardless of our relative international rank, informed citizenship in the year 2000 requires that all people have a substantially greater understanding of science. Recall this fall's ballot initiatives in several states, or consider the supermarket dilemma—"Paper or plastic?" Increasingly we are confronted with questions for which scientific information and ways of thinking are necessary for informed decision making.

Finally, a well-kept secret: science is one avenue through which humans can seek understanding of our place in the universe. The personal fulfillment and excitement that science has to offer benefit everyone. For these reasons, scientists and science educators are taking advantage of the current attention on national education goals to do a better job this time around.

> We need national standards to highlight and promote the best practices of the heroes who do what needs to be done despite the norms.

Science for All Americans is not only the name of an influential book issued by Project 2061—the far-reaching effort of the American Association for the Advancement of Science (AAAS)—but also its goal (Rutherford and Ahlgren 1989). Project 2061 was initiated in 1985, a year in which Halley's Comet came close to the earth, and named for the year in which the comet will return. The project delineates the science that people whose lives span those years will need to achieve scientific literacy. Taking the long view, it defines "science" broadly, to include the natural sciences, as well as the social sciences, mathematics, and technology. Project 2061 is taking a decade to produce curriculum models and blueprints for teacher education, assessment, and other systems that need to change to realize the vision. This slow, deliberate pace respects the premise that you cannot create an airplane by adding wings to an automobile.

In contrast, the National Science Teachers Association (NSTA) has crafted a more immediate solution. The 1892 Committee of Ten recommendation "that it is better to study one subject as well as possible during the whole year than to study two or more superficially during the same time" has led to the high school courses we have today. There is insufficient time for the introduction of concepts, starting with an experiential base and leading later to abstraction and formalization. The learning process has been compressed into too short a time frame for meaningful understanding to occur. To improve this situation, NSTA's Scope, Sequence, and Coordination Project (SSC) recommends that biology, chemistry, physics, and earth science (a significant omission of the Committee of Ten) be taught each year, starting in the 6th or 7th grade and continuing through 12th grade (Aldridge 1992). The SSC slogan, "Every Student, Every Science, Every Year," is a short-term version of what Project 2061 envisions.

Why Standards?

With a far-reaching vision for the future and short-term solutions being implemented in schools today, why do we need national standards for science education? First, standards are criteria by which judgments can be made. They need to be based on a vision, to be sure. But they must also address characteristics of curriculum design so that local educators can select what they want their students to learn: one of Project 2061's curriculum models; a particular version of Scope, Sequence, and Coordination; or another option.

What is needed from the national level is guidance for making those decisions. To understand the need for direction from the national level, one need only watch a mathematics teacher arguing for a better system of assessment on the basis of *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics 1989). Or talk with a textbook publisher who wants to make wise selections among the plethora of material in current K-12 science curriculums. The banner put forward by NCTM's *Stan-*

dards enables everyone to move in the same direction, assured that the risks they take to improve mathematics education will be supported by policies and practices throughout the system.

Of course, good things are happening in science classrooms today, even without national standards—but they happen because of the heroes who do what needs to be done despite the norms. Many generous elementary teachers, for example, spend their own money on science supplies because they know that their students learn best by investigating. Middle schools often use science programs with relevance to their students' lives instead of being merely a practice for high school. And some high school teachers, ignoring the vocabulary-dense syllabus, encourage student inquiry into questions of their own.

We need national standards to highlight and promote the best practices of these heroes. We must make their curriculums the exemplars—the core of teacher preparation programs, the models for instructional materials and assessments, and the basis on which science programs are judged. We need to recognize and encourage the school principals who find money in their budgets for field trips, the parents whose bake sale proceeds purchase science equipment, the authors who write materials that cannot possibly satisfy the divergent criteria of 22 states and thousands of local districts, and the publishers who are pioneering in authentic assessments despite the lucrative market for multiple-choice tests. These leadership efforts must become the goal toward which others strive.

How Are Standards Being Developed?

The National Research Council—the operating arm of the National Academy of Sciences and the National Academy of Engineering—agreed to take the lead in developing standards for science education, at the request of the National Science Teachers Association and other professional societies, the Secretary of Education, the Assistant Director for Education and Human Resources of the National Science Foundation, and the Governors who co-chair the National Education Goals Panel.

While scientists and science teachers are prominent in the process, teachers are a plurality on all committees and working groups. Also involved are other educators, business representatives, and the public. To further rectify the sins of omission of the 1892 committee, women and members of other groups currently underrepresented in the sciences (people of African-American, Latino, and Native American origin) are participating directly in the process. In a further attempt to ensure that the views of 18, or even 180 people, do not set the agenda for standards intended for all students, a plan for broad critique and consensus has been built into the process.

Meeting for the first time in May 1992, the National Committee on Science Education Standards and Assessment approved a plan to produce science education standards by fall 1994. The committee's charge to its working groups on curriculum, teaching, and assessment is:

> ...to develop, in cooperation with the larger science, science education, and education communities, standards for school science.
>
> The standards, founded in exemplary practice and contemporary views of science, society, and schooling, will provide a vision of excellence to guide the science education system in productive and socially responsible ways. Standards for curriculum, teaching, and assessment will be integrated in a single document. The standards will specify criteria to judge the quality of school science and to guide the future development of the science education enterprise.

What Standards Are Being Developed?

Science curriculum, science teaching, and science assessment standards are being created. *Curriculum* standards will define:

- the nature of school science experiences that exemplary practice and learning research propose are effective in producing valued science learning;
- the scientific information (facts, concepts, laws, theories), modes of reasoning, and proficiency in conducting scientific investigations that all students are expected to attain as a result of the experiences;
- the attitudes and inclinations to apply scientific principles and ways of thinking outside the formal educational system that all students are expected to attain.

Science *assessment* standards will define:

- the methods for assessing and analyzing student achievement and the opportunities that programs afford students to achieve the valued outcomes of science;
- the methods for achieving appropriate correspondence between assessment data and the purposes that the data will serve;
- the characteristics of valid, reliable science assessment data and appropriate methods for collecting them.

Science *teaching* standards will define:

- the skills and knowledge teachers need to provide students with school experiences to achieve the valued science learning outcomes;
- the preparation and professional development teachers need to fulfill their roles;
- the necessary support systems and resources for effective science teaching.

The National Science Education Standards will be descriptive, not prescriptive, in order to support thoughtful consideration and application. The curriculum standards will not prescribe particular courses, programs of study, or textbooks; assessment standards will not be an examination; and teaching standards will not be certification or licensure specifications. In each case, examples will illustrate the broad range of what

Prototype Activity: Matter (1st Grade Level)

The following is an example of how content standards can be taught and evaluated. This passage is excerpted from: National Committee on Science Education Standards and Assessment, (December 1992), National Science Education Standards: A Sampler *(Washington, D.C.: National Research Council), pp. 30-33.*

...Today, the teacher had planned to take the class for a walk around the block ... [to] collect rocks for study.... The teacher told the students about the purpose of the walk and asked them what they thought they might find and where they might find them. Divided into pairs and equipped with a map of the block ... and a bag, the children circled the block, stopping to collect stones as they went. Back in the classroom, the students, in groups of four, examined their rocks closely, using hand-held magnifying lenses. They were asked to think about how they might describe their rocks, making drawings if they wished, and then to sort their rocks into groups that made sense to them.

Within their groups, the students discussed their observations and agreed and disagreed about categories. The teacher moved from group to group, listening to the discussions, asking for descriptions, pointing out interesting features, and querying the reasons for the groupings....

The next day, the teacher [asked] each group ... to explain the basis for the grouping of their rocks. Other students were asked to comment. The teacher picked up a new rock and asked that it be placed in the proper pile. After each group of four had completed its explanation, the teacher and the class constructed a list of all the characteristics ... used in sorting the rocks. They discussed the relative usefulness of some versus others and talked about other tools that might be useful....

...This unit will continue. The students will pursue the study of rocks as well as other parts of the environment. In the process, they will continue to study the properties and characteristics of objects and materials and apply their abilities to observe, describe, and classify....

... As a result of [many] activities [like these] students should be able to demonstrate their understanding of fundamental ideas about objects and materials; namely, that:

- Common objects have observable properties (size, shape, volume, and weight) that can be compared and measured ... [and] used to describe, group, and classify objects.

In demonstrating their understanding of these ideas, students should be able to classify or order a set of objects according to a specified property, such as weight or volume.... and to devise one or more ways to classify or order a set of objects and ... to explain their classification scheme.

- Objects are made up of different kinds of materials. Materials have observable properties (color, texture, magnetic characteristics, and different behaviors when heated or cooled) that can be compared and measured. Such properties are useful in describing, grouping, and classifying materials.

...Students should be able to group a set of objects according to the materials from which the objects were made (wood, metal, glass, and clay).... [and] describe differences in the observable properties of such materials.

- Materials can exist in different states (solid, liquid, gaseous). Each state has characteristic properties.

...Students should be able to describe observable properties that given materials have in common or that distinguish them from one another.

- Some properties of a material may change when it experiences external change; others do not. In particular, if the temperature of a sample of materials is changed, the material may change from one state to another (liquid to solid, liquid to gas, and so on). However, the weight of an object remains unchanged when it is broken into smaller parts.

...Students should be able to predict and describe the effects of temperature changes on water or ice.... [and] to provide evidence that the weight of a sample of material remains the same even though its shape, location, or appearance may change.

is possible, not define the one "best" approach.

While working groups in curriculum, teaching, and assessment convened separately in the summer of 1992, their overlapping membership and common goal will lead to a unified document that will move the system of science education forward in concert. The first discussion document, prepared in October,[1] outlined guiding principles for the production of a complete draft by fall 1994. These principles delineate the territory of school science—somewhat broader than Scope, Sequence, and Coordination, but narrower than Project 2061.

Science for All

The first principle, Science for All, takes an unwavering stand that "the science standards will define the level of understanding that all students—regardless of background, future aspirations, or interest in science—should develop." The text foreshadows what will appear in subsequent drafts:

> ... the commitment to "Science for All" implies inclusion not only of those who traditionally have received encouragement and opportunity to pursue science, but of women and girls, all racial and ethnic groups, the physically and educationally challenged, and those with limited English proficiency. Further, it implies attention to various styles of learning and differing sources of motivation. Every person must be brought into and given access to the ongoing conversation of science.
>
> Thus, the commitment to "Science for All" requires curriculum, teaching, and assessment standards that take into account student diversity vis-a-vis interests, motivation, experience, and ways of coming to understand science. The standards must define criteria for high-quality science experiences that include the engagement of all students in the full range of science content. These experiences must teach the nature and process of science as well as the subject matter and support the notion that men and women of diverse backgrounds engage and participate in science and that all have a claim on this common human heritage.

The commitment to "Science for All" has implications for program design and resource allocation at local, state, and national levels.

The other guiding principles in the October discussion document delineate the territory of school science, distinguishing it from technology and engineering and from other ways of knowing. The position taken is that the national science education standards should be limited to the fundamental understandings and should offer selection criteria that states, localities, teachers, and students can use to determine additional subject matter to be studied. Such criteria will include: developmental appropriateness, experiential connections, contribution to students' ability to investigate and to make decisions, and being worth the instructional time and student effort to achieve understanding.

These positions are elaborated in a December discussion document, which provides one or more prototype standards illustrating the interweaving of curriculum, teaching, and assessment. This document also characterizes the domain of science education standards, indicating the inclusion not only of the subject matter (ecology, energy, space, and so on) but also inquiry, decision making, and content (social, ethical, and historical).

Critique and Consensus

Parallel to the development of science education standards is a broad-based critique and consensus process. By working with the several science and science education communities (biology, chemistry, physics, earth and space sciences), we hope to overcome the fragmentation and territoriality that have characterized much of science education in the past. By involving those who have been left out—females (Association for Women in Science); members of racial and ethnic groups (American Indian Science and Engineering Society, Hispanic Secretariat, National Association of Black School Educators); and the physically challenged (Foundation for Science and the Handicapped)—we will work to enlarge the mainstream of the science and science education community.

We distributed the December draft to a wider audience than the October document. The overwhelming sentiment among the more than 5,000 scientists, science educators, and educators with whom we have talked is concern that the public will not support the changes the standards will call for. The publishers and producers of instructional materials and tests, with whom we have had two meetings each already, are also eager to see public support for hands-on, "minds-on" science programs. Similarly, the corporate community is willing to work for standards-based systemic change—and is poised to assist with expertise and funding. The National Governors' Association has put standards-based, systemic reform at the top of its 1992-93 agenda, with mathematics and science leading the way. Other policy groups have given us similar endorsements.

By extending the discussion to include the broader education, business, parent, and policy communities,[2] we hope to create a context of wide support for the science education goals. And by working with the public and private funders of education,[3] we hope to provide the support that teachers will need to meet the standards.

In other words, after 100 years of observation and experimentation, this time we hope to get it right! ■

[1]The document was prepared by working group chairs Audrey Champagne, Henry Heikkinen, and Karen Worth.

[2]For example, Association for Supervision and Curriculum Development, Council of Chief State School Officers, Corporate Council for Mathematics and Science Education, National Parent Teachers Association, and National Governors' Association.

[3]For example, Council on Foundations, National Council of State Legislatures, National Science Foundation, and the U.S. Congress.

References

Aldridge, B. (1992). *Scope, Sequence, and Coordination of Secondary School Science: The Content Core*. Washington, D.C.: The National Science Teachers Association.

Blank, R., and M. Dalkilic. (1990). *State Indicators of Science and Mathematics Education*. Washington, D.C.: CCSSO, State Education Assessment Center.

National Council of Teachers of Mathematics, Commission on Standards for School Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, Va.: NCTM.

National Education Association (1894). *Report of the Committee of Ten on Secondary School Studies*. Chicago: American Book Company.

Rutherford, J.F. and A. Ahlgren. (1989). *Science for All Americans*. New York: Oxford University Press, Inc.

U.S. Department of Labor. (January 1992). *Employment and Earnings*. Washington, D.C.

Weiss, I.R. (1989). *Science and Mathematics Education Briefing Book*. Chapel Hill, N.C.: Horizon Research, Inc.

Authors' note: Project start-up funds of $500,000 and $2.5 million for science education curriculum standards were provided by the U.S. Department of Education. Funding for teaching and assessment standards development and the critique and consensus process to be provided by the National Science Foundation and a coalition of federal agencies: National Aeronatics and Space Administration, National Institutes of Health, and U.S. Departments of Agriculture, Defense, and Energy.

For further information or to receive a copy of the recently revised science standards description and progress report, please fax your request to (202) 334-3159, or mail your request to Ginny Van Horne, NCSESA Office of Critique and Consensus, National Research Council, 2101 Constitution Ave., HA 486, Washington, DC 20418.

Kenneth M. Hoffman is Professor of Mathematics, Massachusetts Institute of Technology, and Associate Executive Officer for Education, National Research Council, 2101 Constitution Ave., Washington, DC 20418. **Elizabeth K. Stage** is Executive Director, California Science Project, University of California Office of the President; and Director of Critique and Consensus, National Committee on Science Education Standards and Assessment, National Research Council.

BEST COPY AVAILABLE

Reprinted with permission from Association for Supervision and Curriculum Development, excerpt from Expanding Student Assessment, edited by Vito Perrone, copyright 1991.

7

Active Assessment for Active Science

George E. Hein

The Need for Active Assessment Methods

Science is an active process that involves using physical skills, imagination, and creativity to tackle the usually ill-defined problems and events of the real world. In looking at our methods for assessing science learning in schools, however, we might think that what's most important in science is being able to choose the one correct answer for each question on a multiple-choice test. Assessing science through multiple-choice tests is like assessing Larry Bird's basketball skills by asking him to respond to a set of multiple-choice questions. We might find out something about Bird's knowledge of the facts of basketball, perhaps even something about his conceptual knowledge, but we certainly would not be able to measure the level of his playing skill.

Increasingly, commentators on the state of science education recognize this mismatch. The National Science Foundation (NSF) in 1987 launched a major curriculum development initiative that began with elementary school projects and will add middle school and high school projects in later years. In a memorandum on assessment written in the second year of this effort, a group of the NSF-supported curriculum developers concluded:

> Research shows that extant achievement tests do not measure the broad range of scientific processes or higher order thinking skills, nor do they give insight into naive versus "scientific" interpretations of phenomena. All these domains are integral to current approaches to teaching and learning science. . . . On the contrary, because the emphasis of these norm-referenced tests is on types of questions that can be answered by

Author's note: I am grateful for the support of the National Science Foundation, Grant #TPE-885032, which contributed to the collection of material for this chapter.

> simple recall of facts, and/or recognition of textbook experiments, they militate against the less predictable hands-on approach. . . . The existing norm-referenced tests not only fail to support or encourage the implementations of new developments in science curriculum and pedagogy, but their continued, near-universal use may dampen or totally inhibit implementation of such approaches. Thus, there is a need for alternatives to existing national, normreferenced tests. The alternatives must be of high quality and must meet the public's needs for accountability and comparability across programs and districts. Additionally, they must be congruent with the philosophy of science teaching and learning the National Science Foundation promotes (Harmon et al. 1988).

The National Center for Improving Science Education, a policy group whose mission is "to promote changes in state and local policies and practices in the science curriculum, science teaching, and assessment of student learning in science" (Raizen et al. 1989), has begun to issue a series of reports covering assessment, curriculum, and teacher training at the elementary, middle, and secondary school levels. In the first report on assessment, the expert panel convened by the Center stresses the need for new and more varied assessment methods and argues for a national program to improve science assessment:

> Improvement Goal 2. Development of externally mandated assessments as well as classroom tests that conform closely to the characteristics of good science curricula and instruction. . . . Assessments should provide greater opportunities for children to interact with stimulus materials, (2) attend to understandings of constructs and principles as well as factual knowledge, (3) probe approaches to problem solving as well as outcomes, (4) be explicitly integrated with the curriculum and with instruction, (5) incorporate hands-on activities whenever feasible, and (6) be structured around group as well as individual activities (Raizen et al. 1989, p. 97).

These are but two examples of calls for a reform, indeed a revolution, in how we assess knowledge of science. Every major policy paper of the past few years, whether focused on national indicators (Knapp et al. 1987, Murnane and Raizen 1988) or on classroom practice (Resnick 1987; Champagne, Lovitts, and Calinger 1990), has called for a similar change in assessment.

Fortunately, developing alternatives to multiple-choice assessments need not start de novo. As long as teachers have wondered about what students have learned, a wide range of assessment strategies and practices have flourished. The dominant use of paper-and-pencil tests at the

national level has only obscured, not eliminated, the alternative work that has taken place in a variety of settings and at levels ranging from the classroom to national-scale assessments. Much recent work is well documented and relevant to any effort to develop classroom-based and large-scale alternative assessments.

In this chapter, I summarize a number of different ways by which learning in science has been and can be effectively assessed, and I describe a few cases in detail as illustrations of more widespread practice. I begin by outlining the various methods that are available to assess student learning in science. I next examine several categories of research and development used to look at learning in science, and discuss the methods professionals in these fields have employed. Finally, I discuss a few issues that emerge from this catalog of methods.

I do not cover a number of technical issues related to assessment. For example, all types of assessment are subject to questions concerning reliability and validity. In general, the simpler the method to administer and score, and the less the method is subject to variation because of local circumstances or context, the easier it is to establish reliability. However, the same conditions generally make the validity of the results more difficult to achieve, since the requirements for a simple, all-purpose test that can be administered in any context usually mean that the assessment differs from the actual activity that is being assessed. Thus, a paper-and-pencil test for science achievement can be made highly reliable, but still leave serious questions about its validity, as the basketball example suggested.

All assessment methods carry with them issues concerning practicality and cost. The cheapest and most practical test, especially for large-scale testing, is one that can be administered to a large group of students simultaneously in minimum time using the fewest materials. But again, the closer a test comes to this ideal, the more likely it is that its validity may come into question.

The most appropriate assessment method for any particular application may also vary with the purpose of the assessment. Generalizable, group-administered, context-invariant assessments, because they are relatively inexpensive, easy to administer, and easy to understand and interpret, are often considered more suitable for large-scale assessments for policy purposes. Individualized, longer, and more curriculum-embedded assessments are considered primarily for their value to the classroom teacher, because they are more complex and the results are usually used for diagnostic purposes. But because

assessments for policy purposes need not involve every child, and may provide valid and reliable information on small population samples, there is little need for them to be simple, and the information lost by making them too generalizable may be greater than what is gained by the simplicity. My goal is not to elaborate further these arguments for or against the various types of methods on the basis of concerns such as reliability and validity, practicality and cost, or purpose of assessment. Instead, I lay out the methods that have been used effectively and illustrate them with examples from assessment contexts and other science education activities in which they have been used. My sympathy is with the use of a wide range of methods that come as close as possible to actual practice (Hein 1987, 1990).

A Survey of Assessment Methods

Observation

Observation is the oldest known scientific method for the study of nature. It was established long before science became a separate form of inquiry and is a common assessment tool for a wide variety of learning activities. In sports such as diving and gymnastics, in music competitions, and in crafts programs, observing what the learner does is a traditional as well as modern way of evaluating achievement. Complex types of learning, such as those just mentioned, and more mundane skills, such as using a measuring instrument or carrying out a filtration, are served equally well.

Psychologists from Itard to Piaget to the present have watched children and adults perform to determine their level of understanding or stage of development. Piaget (1929) argued that the only advantage his chosen method of clinical interviews had over observation was that it allowed the experimenter to contrive situations that might not occur as readily if children's behavior were simply observed. Otherwise, he said, observation would be an excellent research tool.

Duckworth's (1978) assessment of the African Primary Science Project is an example of the use of observation as the primary assessment tool. This curriculum project endeavored to introduce African elementary school children to science through materials-based, hands-on activities derived from a similar curriculum developed at the Elementary Science Study. Children observed the behavior of ant lions (an indigenous insect), worked with simple electricity, used

classroom-constructed microscopes, and so on. What was the value of this program? Did the children who participated in the program learn anything that other children did not?

To answer these questions as an outside program evaluator, Duckworth set up a mock classroom that contained materials similar to, but different from, the ones used in the project. She divided the children into two groups, those who had participated in the project and those who had not. Then she observed each group's behavior in the mock classroom. What she saw was that students who had been part of the project interacted more with the science materials and used them in more complex ways than did the children who had followed the more textbook-oriented curriculum. She was even able to develop a rough quantitative scale of diversity/complexity to compare the work of the two groups. Her evaluation may have been helped by the fact that it was carried out in a culture that did not have an abundant supply of the kinds of materials that are common to hands-on science programs.

Several more recent science assessment schemes involve trained assessors observing what students do as they go about "doing" science. The Massachusetts State Department of Education devised performance tests for a stratified random sample of 3,000 4th and 8th grade students, which teachers administered in the spring of 1989. Students were asked to classify groups of objects, estimate the number of grains of popcorn in a container, and complete various measurement tasks. The teachers who acted as assessors were trained to observe the children's activities and write up what they observed.

These exercises were derived from the major British national assessment effort carried out by the Assessment of Performance Unit (APU), which I discuss in more detail below. Observing children doing science was an important component of the assessment, and the value of observations, as well as the difficulties associated with them, have been discussed at length in the technical and summary reports prepared as part of the APU work. For example, Harlen, Black, and Johnson (1981) describe an exercise in which 11-year-old children were given a mechanical caterpillar that crawled forward when wound up. After the children played with the toy for two or three minutes, they were asked whether they saw a connection between the number of times the wind-up key was turned and the distance traveled by the caterpillar. Trained testers watched the children and recorded each child's behavior on a prepared checklist. The APU group found observation difficult, but possible, given

time to train the testers, forethought concerning the possible activities that children might undertake, and one-to-one test administration:

> In some cases it was not possible to decide whether, for instance, an action which apparently controlled a variable was deliberate or accidental, and in such cases the decision had to be left until later when the pupil's work was discussed with him. Using the check list was a skilled activity requiring intense concentration on the part of the tester, for a wavering of attention might result in an action being missed, with no opportunity to replay or recall the event. It was often a struggle to make sense of the pupil's action (what do you do when a pupil uses the string to tie the caterpillar's hat to the leg of a chair and proceeds to get the toy to tow the chair?) and it was the need for this close scrutiny that made one-to-one administration essential (Harlen, Black, and Johnson 1981, p. 115).

Verbal Responses

Verbal responses are a particularly useful way of finding out what students know, since they make up much of the day-to-day interchange between teachers and pupils. As more formal assessment methods, they also have their place. In the quotation above, for example, it is evident that testers talked with children and asked them why they carried out the actions they did. Researchers interested in children's concepts often conduct clinical interviews to investigate children's knowledge of and ideas about science. In advanced degree work in all fields, oral examinations in which candidates and professors engage in discussion to find out what the student knows are standard practice.

Churchill and Petner (1977) suggested that children's spontaneous conversations can be a guide to their science knowledge. And Chittenden (1990) has studied the idea of group conversations as a basis of classroom assessment. He suggests that teachers carry on conversations with an entire classroom of students, following a few guidelines:

- that discussion begin with open-ended questions, such as:
  - –What have you noticed lately about our caterpillars?
  - –What are some things you know about shadows? What is a shadow?
  - –What sorts of questions do you have about the sun? What have you wondered about?
- that teachers refrain from correcting or unduly modifying the children's comments

- that discussions proceed in a manner ensuring the involvement of most all of the children
- that records be made of each child's statements (Chittenden 1990, p. 221).

The guidelines are intended to encourage the sort of discussion that is sustained by child-initiated questions and ideas, and that allows children some control over the direction of the conversation.

Chittenden characterizes the discussions as "staged observations" that attempt to capture a dimension of classroom life that ordinarily remains undocumented—namely children's talk. Dyasi (1990) has also analyzed transcripts of teacher-student conversations to gain an understanding of children's approaches to science. He sees the recorded discussions as an important component of the documentation of students' learning, part of a portfolio describing students' progress.

Formal recording of conversations requires considerable resources. If observation requires one-to-one interaction, then guiding and recording conversations requires more than a one-to-one ratio of assessors to groups of students. Although conversations can be recorded on tape, they still need to be listened to, possibly transcribed, and analyzed. Nevertheless, children's verbal responses can be made part of classroom assessment. Even if carried out less formally than would be required for research purposes, they can provide the teacher with information on what children know and understand, and notes from such conversations can be part of an assessment system. Several curriculum development projects suggest that teachers talk with students both before a science unit is introduced and after it is completed. The first conversation helps teachers assess students' baseline knowledge, and the second conversation gives them the opportunity to determine how students' knowledge has changed.

Written Records

Written responses from students are a simple and direct way to assess learning; in fact, they are the most common form of academic assessment. Our reliance on multiple-choice and other short-answer tests, however, may make us forget that written tests can take many forms, from completing sentences, arranging statements, and making lists to providing explanations for answers, composing essays, and drafting reports. Carlson (1987) provides examples of ten common types of written examinations, including many specific examples from secondary science education.

The General Certificate of Secondary Education (GCSE), which was introduced in all British state-supported secondary schools in 1988, uses a variety of short and long written response methods, as well as practical tests, to assess learning (Department of Education and Science 1985).

Written work can reflect students' science knowledge without being specifically part of an assessment task. Thus, Rockcastle (1986) argued that essays may provide considerable insight into students' science knowledge. He uses as an example stories children wrote in response to the prompt to describe what their school might look like more than 100 years from now, considering the effects of biological succession. More test-like are the explanations children give for their answers to test questions. A common type of question used on APU tests and GCSE examinations asks students not only to give an answer, but also to provide one or more reasons for their answer. An entire category of the APU scheme concerns planning investigations. Students are presented with a problem, provided with a written list or pictures of material they might use to solve the problem, and asked how they would go about doing so (see Figure 7.1).

FIGURE 7.1

Suppose you are going to make a chopping board to use for cutting bread or chopping vegetables or meat. You have to decide which is the best kind of wood to use. You have blocks of four different kinds of wood (A, B, C, D) and you can use any of the things in the picture below to do some tests on them. (You don't have to use all the things).

What would you do to:

Test the blocks to find out which kind of wood is best for making a chopping board.

Make sure you say (1) which things you would use (2)what you would do
(3) how you would find out the result

A number of APU practical tasks requiring written responses were adapted by the National Assessment of Educational Progress (NAEP) and tried out in the United States (Blomberg et al. 1986). NAEP (1987) published a popular version of these methods for teachers.

Written responses can also be the result of actual science work. Students can be asked to carry out various tasks that will result in answers to given questions; if students perform the tasks correctly, they should come up with the correct written answers. In several assessments of measurement skills, including the ones carried out in New York State and Massachusetts and by the APU, students are given a thermometer or other scale to read, or a ruler and an item to measure, and then asked questions such as "How long is this item?" or "What is the reading on this scale?" The written test paper can be used as evidence of the student's degree of success at the task.

Drawing

Through illustration, students can demonstrate an idea or concept or show that they have learned a skill. They may sketch what a product looks like, describe an apparatus by detailed drawings or diagrams, depict a situation, or illustrate their beliefs. As an assessment tool, drawing can range from the most artistic, free expression to the precise rendering of technical details. Dyasi (1990) has analyzed drawings that are part of the portfolios of students' work from the Prospect Archive and discussed how they can be used to provide information about students' knowledge of science. In my own work (Hein 1985), I found that teachers could test students' knowledge of how to use a microscope simply by asking them to draw what they saw through the microscope (see Figure 7.2).

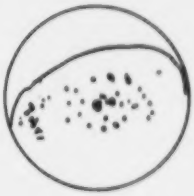
Products

Practical work in science often leads to products. And examining the products of students' practical work can indicate what students have learned. If an animal is cared for, if a doll house is wired and the lights work (a final assessment task for a curriculum unit on electricity), if the product of the chemical reaction is crystalline and pure, we can make inferences about a student's level of performance and understanding.

The APU surveys used products in an ingenious way. Students at each of the three age levels were asked to carry out multistep procedures that resulted in a product. The purpose of one such activity, building a simple kaleidoscope from folded paper, tape, and mirrors, was not to

FIGURE 7.2
Student Drawings

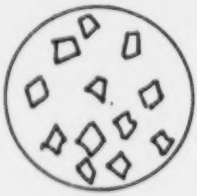
Part Two
Draw a picture in the circle below of what sand looks like through a microscope.



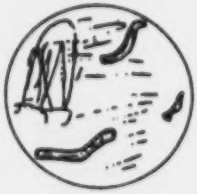
Part Three
Draw a picture in the circle below of what you would see in a drop of water taken from a pond, when you looked at it under a microscope.

Pretest
MOS Kits Evalution1986

Part Two
Draw a picture in the circle below of what sand looks like through a microscope.



Part Three
Draw a picture in the circle below of what you would see in a drop of water taken from a pond, when you looked at it under a microscope.



Post-test
MOS Kits Evalution1986

produce the product, but to determine how well students could follow instructions. In other assessment schemes, such as those developed in Connecticut (Baron 1989), students are asked, individually or in groups, to carry out both short-term (less than one period) and long-term tasks (over several days to several weeks) that result in a product so that teachers can evaluate students' knowledge.

In summary, we can assess students in a variety of ways: we can observe what they do, listen to what they say, read what write, and analyze what they produce. Any behavior that can be perceived can be adapted for assessment. The typical written, short-answer test is just one point on the continuum of assessment.

An Assessment System: The APU

Much of the literature critical of present science assessment practices argues not only for *alternatives* to multiple-choice tests, but also for a *variety* of methods to assess student performance. The most extensive model for such an alternative assessment for large-scale national policy purposes is provided by the APU (Black 1987), which systematically collected data on British student achievement for over a decade. Established by the Department of Education and Science in 1975 "to promote the development of methods of assessing and monitoring the achievement of children in schools and to seek to identify the incidence of underachievement" (quoted in Harlen et al. 1981), the APU conducted national achievement surveys in certain school subjects. The APU science monitoring teams began their work in 1977 and developed an assessment framework based on the proposition that "science is to be regarded as a mode of thought and activity which may be encountered in a number of subjects appearing in the school subject."

In five annual surveys of schools from 1980 to 1984, the APU collected 3,750,000 responses from 240,000 pupils aged 11, 13, and 15 in 7,500 schools across England, Wales, and Northern Ireland (a 2 percent sample). The surveys provide a fairly detailed description of the level of science competence of British students, and the methodologies and outcomes of the surveys have profoundly influenced the new National Curriculum, which was introduced into all Welsh and English state schools in fall 1989. The APU work has provided insight into students' understanding of concepts, pioneered evaluation methodologies, and spawned major research programs based on both the findings and the methodology of the surveys.

In designing the overall plan to survey science achievement, the APU tried to take into account that science is primarily a way of doing things, and only partially a collection of facts and concepts. The group set up a six-part scheme to describe science and then developed different kinds of assessment strategies for each component (see Figure 7.3).

FIGURE 7.3
The Categories of Science Performance

1. Use of graphical and symbolic representation	— reading information from graphs, tables, and charts	written test
	— representing information as graphs, tables, and charts	
2. Use of apparatus and measuring instruments	— using measuring instruments	group practical test
	— estimating physical quantities	
	— following instructions for practical work	
3. Observation	— making and interpreting observations	group practical test
4. Interpretation and application	— i interpreting presented information	written test
	— ii applying: Biology concepts Physics concepts Chemistry concepts	
5. Planning of investigations	— planning parts of investigations	written test
	— planning entire investigations	
6. Performance of investigations	— performing entire investigations	individual practical test

In the APU scheme, assessment of science knowledge and concepts is limited to categories 4 (Interpretation and application) and 5 (Planning of investigations); the other categories focus on processes. The actual interdependence of concepts and processes (as well as a third component, attitude or interest) was recognized by the group at the

BEST COPY AVAILABLE

beginning of the assessment and repeatedly reinforced by the results. Nevertheless, to the extent that the categories can be separated, the APU approach does so.

The practical tests for category 2 were usually administered to groups of students at stations set up in a classroom. Students went from station to station and carried out the measurements and other tasks as directed. Category 6 tasks were administered one to one. The wooden board example in Figure 7.1 was used as both a practical test and a planning exercise, as were dozens of other tasks. Students were given materials and equipment and asked to demonstrate which piece of wood made the best cutting board. The written sections varied with the different categories and included many alternatives to short-answer questions. Figures 7.4 and 7.5 illustrate other types of questions that were used in this assessment.

Assessment Methods in Research and Curriculum Development

If we wish to probe children's science beliefs, or understand how concepts develop, we have to find out what children know. Similarly, in order to assess the value of any science curriculum, we have to find out what students learn from using the curriculum. One way to gauge the adequacy of the current state of science assessment is to examine the extent to which curriculum developers and researchers interested in children's understanding of science employ current tests in their work. The curriculum groups provide a particularly appropriate touchstone, since their goal is to introduce new materials into existing schools. We can also look at what assessment tools the curriculum developers actually employ as they produce science materials and introduce these materials into the classroom.

In general, newer science curriculum materials advocate assessments that:

- are embedded within instructional materials,
- use a variety of methods to assess the student's progress,
- emphasize teacher observation and teacher judgment,
- provide methods for getting at the reasons behind children's answers.

For example, The Improving Urban Elementary Science (IUES) project (Harmon and Mokros 1990) uses a general assessment framework for all its units. At the start of instruction, the teacher gives students a

FIGURE 7.4
APU Sample Questions

Category 5: Planning of Investigations (Age 13)

A group of pupils are comparing water from two different towns. They want to do a test to find out which kind of water lathers more easily with soap flakes.

If they want to make it a *fair* test they will have to make sure that some things in the test are the same for both kinds of water. Suggest *three* things that should be the same:

1. _____
   _____
2. _____
   _____
3. _____
   _____

Category 4: Interpretation and Application (Age 11)

Walking along this footpath, Thomas noticed that there was ivy growing on the trees, but only around three-quarters of the trunks. None of the trees had ivy growing on the side nearest to the path.

Think of *two* different reasons why the ivy might grown only on some sides of the trees. Write the first under (a) and the second under (b).

(a) I think it might be because _____
_____

(b) I think it might be because _____
_____

pre-unit questionnaire to gather baseline information about their knowledge of the subject. The questionnaire requires writing and drawing as well as short answers. While carrying out the unit, the teacher is advised to observe certain aspects of students' behavior as evidence of learning, and assessment modes are provided through the course of the unit, including embedded assessments. "An embedded assessment is not a test. It is one of the daily learning experiences written in a special format" (EDC 1989). At the end of the unit, both written and performance assessments are provided. In a 6th grade unit on structures, in which children build towers and other architectural objects and examine what makes structures stand, an embedded assessment involves building bridges; the final performance assessment requires students to design and build a model of a playground (EDC 1989). The written items in the assessment sections ask students to draw features of structures and to explain what they understand concepts to mean; they also provide many opportunities for open-ended responses.

Thus, the project materials include two kinds of assessment activities. One category involves tasks that are part of the curriculum and that provide feedback to teachers (and to others) as students progress. Another category constitutes more formal assessment at the end of a set of activities to provide summary information on what students have learned. The written material for the assessment component of the unit emphasizes that a multiplicity of methods is not only desirable but necessary to find out what children have learned during the course of the varied hands-on activities contained in the unit.

A similar multiplicity of assessment methods characterizes the work of other research groups interested in understanding children's science concepts. Some typical methods include the following:

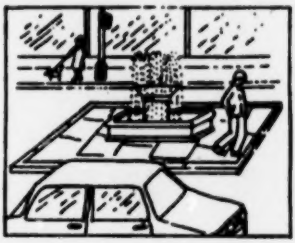
Clinical Interviews: This method, so brilliantly employed by Piaget, is still extensively used. Carey's (1985) insights into children's understanding of living things, obtained primarily through interviews, is a fine example. In some instances children are interviewed without being shown any prompts; in other cases they respond to drawings, photographs, or objects, or to questions about an activity they have carried out.

Drawing: Students' understanding of the nature of light has been explored by a number of research groups using, among other methods, the simple device of asking students to draw what happens when the eye sees an object (Anderson 1983, Chittenden 1984, Osborne et al. 1990).

FIGURE 7.5
APU Sample Questions

Category 4: Applying Chemistry Concepts (Age 13)

A smooth marble fountain was built in the middle of a city.



After several years, the surface of the marble was worn and covered with small holes.

Think of three reasons, other than damage by people, which could have caused the small holes to form.

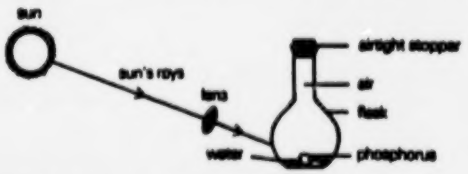
1. _____

2. _____

3. _____

Category 4: Applying Chemistry Concepts (Age 15)

A piece of phosphorous was held in a flask as shown in the diagram. The mass of the flask and contents equalled 205 g. The sun's rays were focused on the phosphorous, which then caught fire. The white smoke produced slowly dissolved in the water.



After cooking, the flask and its contents were weighed again.

1. Would you expect the weight to be
   - __ A. more than 205 g.
   - __ B. 205 g.
   - __ C. less than 205 g.
   - __ D. not enough information to answer

2. Give the reason for your answer:

_____

_____

Performance: In a series of experiments intended to discover young adults' and adults' understanding of physical forces, McClusky (1983) asked individuals to walk across a room and drop an object at the appropriate moment so that it would land in a container. Driver (1990) has summarized the status of research on conceptual development and its relationship to science assessment.

Researchers interested in exploring students' understanding of science usually use assessments that go far beyond the boundaries of traditional tests and they rarely use multiple-choice questions.

Assessment Issues

Comparing Methods

An obvious issue raised by the availability of such a wide range of assessment methods is whether different methods provide similar or different results. Are data resulting from different types of questions comparable? Do performance measures provide the same information as written measures, only in a different form? Much evidence suggests that even small changes in the framing of questions leads to significantly different results. In the first International Educational Assessment, students answered both multiple-choice questions and practical questions in science. The authors concluded:

> Perhaps of special interest, in view of the current debate on the place to be accorded to practical work of various kinds in school science, was the attempt to produce optional tests of practical abilities requiring only very simple and easily obtainable materials. Unfortunately, only two countries elected to take these "practical" tests, but the evidence from these suggests that such practical tests measure quite different abilities from those assessed by more traditional tests, even those designed to assess practical skills as far as possible without resort to actual apparatus (Comber and Keeves 1973).

In 1984-85 a statewide assessment in Connecticut included both performance testing and multiple-choice items. In one item, reproduced in Figure 7.6, the percentage of correct responses was reduced from 71 percent to only 5 percent when 4th grade students were asked to demonstrate how they would hold the coins rather than choose the correct answer from the list of responses (Baron 1986).

FIGURE 7.6
Example of Science Choice Item Measuring Higher-Order Thinking Skill

Exhibit 1

Percentage of Students Selecting Each Option

Gr. 4	Gr. 8	Gr. 11	
17	31	30	A
71	56	57	B
8	8	9	C
4	3	3	D
1	1	0	No response

Suppose that you want to drop a penny and a quarter at exactly the same time and have them hit the floor at exactly the same time. Which picture BEST shows how you would hold the penny and the quarter just before you drop them?

A. B. C.

D. I don't know

Source: Adapted from Joan B. Baron, "Assessing Higher Order Thinking Skills in Connecticut: Lessons from Connecticut," in *Assessing Higher Order Thinking Skills* (ERIC/TME Report #90)

Badger and Thomas (1989) report significant differences in responses to multiple-choice items and open-ended written questions in their analysis of results from Massachusetts. The APU reports that for similar questions, response rates are, from lowest to highest, as follows:

- written response to questions presented in words
- written response with an illustration
- written response with actual equipment to look at
- written response after actual practical work
- observing the students working.

The Model of Science

Using varied types of assessments more adequately reflects the multifaceted nature of science, and it sets a more appropriate example for the kind of science that should be taught in schools. If a state, a school district, or a national agency advocates hands-on science in the curriculum (as most do), assessment methods should reflect the science described in curriculum guidelines. For example, if a major goal of science teaching is to increase skills for solving poorly defined problems, then students should be required to practice these skills and be assessed for their competence in this area

What Is Being Assessed

The use of assessment techniques that actively involve students in doing science gives us the added benefit of beginning to understand the complex factors that influence performance. Multiple-choice questions, no matter how carefully constructed and how extensively field-tested, must of necessity hide a wide range of reasons for different responses. The thinking behind students' answers can only be inferred from the limited data. For this reason, research groups generally avoid multiple-choice questions.

The APU has field-tested thousands of items through a combination of techniques, trying out similar questions in a range of formats. It has concluded that even very subtle differences in illustration, presentation, or content may profoundly change the way in which students respond. Whether questions are set in a "scientific" or "everyday" context can change answer rates and greatly influence gender differences in responses. When a question stem includes material related to a social issue, students' ability to generalize is diminished because the content distracts them from the data (Donnelly and Welford 1989).

The APU has concluded that the number of factors that influence response rates is so great that we cannot predict, even among questions of similar types, which ones will be easier or harder for particular groups of students. In reviewing the research on science concepts, McDermott (1984) found that certain factors, shown in Figure 7.7, have to be considered when carrying out research on conceptual understanding. These same issues are relevant to science assessment.

Relationship to Instruction. The close relationship between assessment strategies and instruction is supported by using a variety of assessment methods. Since good science instruction invariably involves students' active participation in constructing knowledge in collaboration with their teacher, the passive quality of multiple-choice tests disrupts the instructional flow. It is a separate activity, bringing with it the typical qualities of "testing": anxiety, comparison between pupils, and a change in mood and classroom climate. As the curriculum projects discussed above illustrate, the use of a wider range of assessment methods allows testing to be embedded into the curriculum.

A wider range of assessment strategies also allows teachers to better understand each student's level of comprehension or skill attainment. Teachers know that the extra time needed to grade problems, read essays, or assess constructions (in comparison to marking multiple-choice

FIGURE 7.7
Characteristics of Research on Conceptual Understanding

Because results and methods are so closely intertwined in research on conceptual understanding, it is important in interpreting the findings to bear in mind the procedures used. Characteristics that should be considered in interpreting the results of a particular project include:

Nature of instrument used to assess understanding. How actively involved was the student in the task? The responses a student makes in writing answers to printed questions may not be the same as those triggered when the student is observing a demonstration, using a computer, or manipulating apparatus in the laboratory.

Degree of interaction between student and investigator. Was it possible for the investigator to clarify student responses? Through further questioning during an interview, an investigator may verify the meaning of a particular response and follow up on comments indicating unsuspected difficulties. On the other hand, during a written examination a student's responses are unlikely to be influenced by what the investigator does or says.

Depth of probing. In how much detail did the investigator examine student understanding? The investigator's perception of student thinking may differ if only one question is asked about a concept rather than many, or if only one context is used rather than several. Results based solely on a student' initial responses may be different from those obtained when the student has the opportunity to consider alternatives.

Form of data. What kinds of data were obtained—for example, written responses to questions, transcripts of interviews, classroom observations? Administering written questions to large numbers of students is useful in determining the frequency of misconceptions in different populations. In contract, the highly interactive structure of an individual interview allows the investigator to examine in detail the nature of a particular difficulty.

Physical setting. In what ways did the environment in which a study was conducted affect the results? A specially designed experimental setting allows an investigator to focus on a given student's understanding a particular concept. However, observing the interaction among students in the more natural setting of the classroom may provide a broader perspective on the range of student beliefs.

Time frame. At what point in instruction was a particular test administered? Over what period of time was the whole study conducted? The significance of particular results may depend on whether tests were administered before, during, or after instruction. Results based on a single administration of a test may differ from those obtained with more extensive testing.

Goals of investigator. How did the perspective of the investigator affect the design of the study or the way in which the data were interpreted? For those who teach physics, the primary motivation in undertaking this kind of research is often the improvement of instruction. For others, the emphasis may be on developing models of human thought. Similar data may be used by some investigators to identify and describe specific difficulties and by others to infer the conceptual framework within which an individual views the physical world.

questions) is compensated for by the insights gained; teachers find out how students tackle problems. A wider range of methods is also necessary if an assessment system is to be applicable to students of all abilities and from all ethnic and social backgrounds. Repeated studies have demonstrated (Meier 1973, Haney 1978) that all kinds of examination and assessment questions can be misunderstood. Only in formats where the assessor can understand the reasons for the answers students have given can cultural and linguistic misunderstandings be analyzed and valid assessments made of the knowledge of all students (Harmon and Mokros 1990).

Assessment Portfolios. Recognition of the need for a variety of assessment methods has led to proposals that assessment be based on a collection of a student's work, a portfolio of materials, rather than by the "blurred snapshot" provided by a single test (Collins 1990). Portfolios of student work have been a cornerstone of the assessment of student progress at the Prospect School, in North Bennington, Vermont, for 20 years, and have been recommended by such diverse agencies as the Coalition of Essential Schools (Wiggins 1989), the Task Group on Assessment and Testing (Black 1987) as part of the English National Curriculum, the Connecticut State Department of Education (Baron et al. 1989), and a task force considering assessment for the Boston Public Schools (*Boston Globe* 1989). Portfolios or profiles may constitute one component of the British CGSE assessment (Brown 1988) and represent a major part of one approved scheme, in which the students assess much of their work themselves (Davis 1989).

Formative and Summative Assessment. A useful distinction can be made between ongoing assessment, during the course of a semester to assist a teacher in preparing lessons and helping students to learn, and final assessments, usually at the end of a unit or year to find out what has been accomplished. The former are formative and the latter summative assessments. Externally developed and administered tests are often considered more appropriate for summative assessments because they avoid the danger of teacher bias and may provide comparable information for a range of classrooms. However, *cumulative* formative assessments that provide evidence over a longer period of time can be just as objective and comparable across classrooms or districts. These might included samples of students' work from an entire semester, evidence for achievement based on carefully defined criteria, or the portfolios mentioned above.

Relationship to Inservice Training. There is obvious value for inservice education in the kinds of data that result from more extensive children's

responses to science probes. Churchill and Petner (1977), Chittenden (1990), and many others have explicitly made use of such information for inservice work. Much of the APU assessment was carried out by classroom teachers, and the major portion of the assessment that will form one component of the new national curriculum will also be in the hands of classroom teachers.

The Task Group on Assessment and Testing, in its recommendations for carrying out assessment at the national level (Black 1987), has proposed a process of *moderation*, a process by which groups of teachers at various levels—the grade, the school, or the district—get together and discuss results and compare grading standards, especially on the more complex open-ended questions and performance measures. The proposed moderation scheme assures that grading will converge on a uniform set of standards, and it serves as a continuing inservice activity for teachers. Practicing professionals would do more than compare student achievement from school to school; they would also compare their own understandings and standards with those of colleagues.

Conclusions

As we have seen, there are diverse methods available for assessing science learning and a wide range of contexts in which these assessments have been used. Many ways of empirically assessing student learning have been developed and applied directly either to classroom-based or larger-scale assessments. The problem of introducing these methods into schools and school systems on a national scale, however, has clearly not been solved.

Changing school practices in any area is a difficult process. Factors ranging from the ordinary inertia inherent in any system to the particular political forces that come to bear on education make it difficult to bring about change. One necessary condition for change is the demonstrated existence of viable alternative practices. They do exist in the field of science assessment. Other components that are needed for such a change include the following:

1. *Time.* Schools and school systems need to embark on systematic, long-term programs to change the nature of assessment. Some of the strategies proposed by state departments and some school reform groups point in this direction. Current testing based on multiple-choice, short-answer questions is established, teachers teach with it in mind, students expect it, and parents and administrators are used to the form of

the results. In order to make a change, every constituency needs time to get acclimatized to the new models.

2. *Time-out.* Time is not enough; "time-out"—a chance to try out new assessment methods without the pressure of performance and accountability based on the old system—is also important. It is unrealistic to expect any responsible educator to embrace a wholly new form of assessment, with uncertain results, as long as funding, professional advancement and perhaps even job review are based on the outcomes of an older system that is to be replaced. Schools and school systems need trial time, a chance to modify practices without the expectation of immediate success and positive results.

The kinds of assessment discussed in this chapter represent a major change in school practice, so they will cause some disruption before they are established. In Great Britain, the new national assessments that accompany the national curriculum will be phased in gradually, with the first year of national testing (at one age level) in 1991 carried out as unreported results, so the first assessment will not take place until 1992, three years after the first children have entered under the new guidelines.

3. *Education.* If we want to change the assessment methods used in schools, then the entire population involved needs to be educated to accept and implement the changes. A new kind of assessment requires rethinking and refocusing. If teachers are to collect portfolios of work, if principals are to receive and to prepare narrative reports of student progress, if state agencies are to make decisions based on a different kind of evidence, it is not enough to argue that this new system is better, provides more valid information, or will be more useful in the long run. We must also provide workshops, inservice training, and time for all the constituencies to discuss and understand the methods and their implications. Most educators believe they know how to interpret the results from multiple-choice science tests, if for no other reason than they have become so familiar with them, that they can relate the results of the tests with their experience. A similar body of knowledge and common understanding needs to be developed for alternative assessments with the different quality of information these will provide.

4. *Resources.* Change requires resources: teacher education, administrator education, public awareness, and a recognition that assessment is a form of passing judgment and can never be made totally objective. It requires a component of professional judgment to interpret the results.

Responsible assessment is a difficult and delicate process. It constantly faces competition from simplistic methods that appear to be more efficient, but are inadequate for carrying out the same task. To establish and preserve valid assessment practices, educators, politicians, and the public need to make a concerted effort to champion a range of methods. The methods are available, and they provide the kind of information that makes for useful debate and discussion. But debate and discussion are not enough. If we intend to improve the way science is taught—as our national education goals claim we do—we must also improve the way it is assessed: active science demands active assessment.

References

Anderson, C.W., and E.L. Smith. (1986). *Children's Conceptions of Light and Color: Understanding the Role of Unseen Rays*. East Lansing, Mich.: Institute for Research on Teaching, Research Series #166.

Assessment of Performance Unit, Science Project. (1989). *Selected Bibliography of APU Publications, 1985–1989*. London: APU Science, Centre for Educational Studies, Kings College.

Badger, E., and B. Thomas. (1989). *On Their Own: Student Response to Open-Ended Tests in Science*. Quincy, Mass.: Massachusetts Department of Education.

Black, P. (1987). *Report: National Curriculum: Task Group on Assessment and Testing*. London: Department of Education and Science.

Black, P. (1990). "Looking to the Future." Lecture presented at the Association for Science Education meeting, Lancaster, England, January 6.

Baron, J.B. (1986). "Assessing High Order Thinking Skills in Connecticut." In *Assessing High Order Thinking Skills*. Princeton, N.J.: Educational Testing Service.

Baron, J.B., P.D. Forgione, Jr., D.A. Rindone, H. Kruglanski, and B. Davey. (1989). "Towards a New Generation of Student Outcome Measures: Connecticut's *Common Core of Learning* Assessment." Paper presented at AERA Annual Meeting, San Francisco, Calif.

Blomberg, F., M. Epstein, W. McDonald, and I. Mullis. (1986). *A Pilot Study of Higher Order Thinking Skills Assessment Techniques in Science and Mathematics: Final Report, Parts 1 and 2*. Princeton, N.J.: National Assessment of Educational Progress.

Boston Globe. (Dec. 12, 1989). "New Academic Tests for Hub Pupils Proposed."

Brown, P. (1988). "Pupil Profiles." In *Assessment At 16*, edited by K. Selkirk. London: Routledge.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, Mass.: MIT Press.

Carlson, S.B. (1987). *Creative Classroom Testing*. Princeton, N.J.: Educational Testing Service.

Champagne, A.B., B.E. Lovitts, and B.J. Calinger. (1990). *Assessment in the Service of Instruction*. Washington, D.C.: American Association for the Advancement of Science.

Chittenden, E., et al. (1984). "A Pilot Study of Science Assessment." Unpublished manuscript. Princeton, N.J.: Educational Testing Service.

Chittenden, E. (1990). "Young Children's Discussions of Science Topics." In *The Assessment of Hands-on Elementary Science Programs*, edited by G.E. Hein. Grand Forks, N.D.: North Dakota Study Group on Evaluation.

Churchill, E.H.E., and J.H. Petner, Jr. (1977). *Children's Language and Thinking: A Report of Work-in-Progress*. Grand Forks, N.D.: North Dakota Study Group on Evaluation.

Reprinted with permission from TEACHERS' LAB, September 1991.

The nature of elementary science: what does "it" look like?

by Gregg Humphrey

IMAGINE A PARENT, or perhaps the superintendent, or maybe another teacher in your school, asking you the question: "What does the kind of science that you are doing look like? What is IT?" You might be tempted to reply by using words such as hands-on, activity-based, inquiry, or process; but the questioner asks: "If I were a child in your room, what might I be doing on a typical day?"

These are questions that the participating teachers and principals of the Vermont Elementary Science Project (VESP) began asking themselves last fall. These educators have now constructed and articulated an understanding of good science for their classrooms. Their Reference Guide outlines what it is that students would know and be doing if they were engaged in hands-on, inquiry-based science.

This profile of effective student practice is the result of teachers and administrators applying the knowledge and methods of the Vermont Elementary Science Project, and then reflecting upon their classroom experiences with each other. This collaborative view of "ideal" student outcomes is particularly relevant to the development of new programs and to the assessment of inquiry-based, elementary science.

These educators have now constructed and articulated an understanding of good science for their classrooms.

As interest in the question of, "What does IT look like?" grew, we strategized ways for educators to feel ownership in the process of developing profiles of effective science practice. The process used to produce these results is interesting in its own right. We asked teams to brainstorm ideas regarding good elementary science for the student. In other words, "What is IT," from the point of view of the child?

During the next phase of developing a practice profile, a committee met and attempted to categorize the data. Working in pairs, and then with the full committee, it was decided to use seven categories into which the ideas were separated (see page 9). Each committee member was then asked to write a short statement to capture the essence of a given category. These statements were to describe "the ideal."

At the next follow-up day of all VESP participants, the large group was separated into seven teams, one for each category of the "IT."

Next, the teams viewed a videotape of 2nd graders investigating sinking and floating. We posed the question, "Is this IT?" Each group was charged with viewing the tape from the point of view of the category that it was concerned with. Afterwards, each team shared with the large group "indicators" that were found from their particular category. More discussion took place refining the wording and ideas of the overall profile.

Finally, the committee met again and produced the work in progress known as "On the Run Reference Guide to the Nature of Ideal Elementary Science for the Student."

By increasing our ability to communicate about the "IT" with one another, we hope to be better able to plan the next steps: for the child, for the teacher, and for the school. If we continue to develop a shared sense of the "IT," we all will be able to assess our progress towards defined goals.

To use the analogy of a microscope with several objective lenses, there are different levels of magnification which help to define the "IT" for the student, the teacher, the school and, ultimately, the "IT" for the school system. We will delve into each of these layers as the Vermont Elementary Science Project continues its work with teachers.

GREGG HUMPHREY is the Technical Assistance Specialist for the Vermont Elementary Science Project (VESP). The VESP is a three-year grant awarded to the NETWORK, INC., Andover, MA, by the National Science Foundation. The ideas in this article were developed by teams of teachers and administrators from eleven schools in the Champlain Valley region of Vermont. It is their wish that this work be viewed as "in progress" and subject to revision. For more information, contact Maura Carlson, VESP Project Coordinator, Trinity College of Vermont, McAuley Hall, Burlington, VT 05401 (802-658-3664).

BEST COPY AVAILABLE

On the run reference guide to the nature of elementary science for the student

To help answer the question: If students are really "doing" hands-on inquiry based science, what does it look like?

Children view themselves as scientists in the process of learning.

1. They look forward to doing science.
2. They demonstrate a desire to learn more.
3. They seek to collaborate and work cooperatively with their peers.
4. They are confident in doing science; they demonstrate a willingness to modify ideas, take risks, and display healthy skepticism.

Children accept an "invitation to learn" and readily engage in the exploration process.

1. Children exhibit curiosity and ponder observations.
2. They move around selecting and using the materials they need.
3. They take the opportunity and the time to "try out" their own ideas.

Children plan and carry out investigations.

1. Children design a way to try out their ideas, not expecting to be told what to do.
2. They plan ways to verify, extend or discard ideas.
3. They carry out investigations by: handling materials, observing, measuring, and recording data.

Children communicate using a variety of methods.

1. Children express ideas in a variety of ways: journals, reporting out, drawing, graphing, charting, etc.
2. They listen, speak and write about science with parents, teachers and peers.
3. They use the language of the processes of science.
4. They communicate their level of understanding of concepts that they have developed to date.

Children propose explanations and solutions and build a store of concepts.

1. Children offer explanations from a "store" of previous knowledge. (Alternative Frameworks, Gut Dynamics).
2. They use investigations to satisfy their own questions.
3. They sort out information and decide what is important.
4. They are willing to revise explanations as they gain new knowledge.

Children raise questions.

1. Children ask questions (verbally or through actions).
2. They use questions to lead them to investigations that generate further questions or ideas.
3. Children value and enjoy asking questions as an important part of science.

Children use observation.

1. Children observe, as opposed to just looking.
2. They see details, they detect sequences and events; they notice change, similarities and differences, etc.
3. They make connections to previously held ideas.

Work in progress. Vermont Elementary Science Project, Trinity College, McAuley Hall, Burlington, VT 05401 (802) 658-3664.

On The Run
Reference Guide To
The Nature of Elementary Science for the Student

To help answer the question: If students are really "doing" hands-on inquiry based science, what does it look like?

Children View Themselves as Scientists in the Process of Learning.
1. They look forward to doing science.
2. They demonstrate a desire to learn more.
3. The seek to collaborate and work cooperatively with their peers.
4. They are confident in doing science; they demonstrate a willingness to modify ideas, take risks, and display healthy skepticism.

Children Accept an "Invitation to Learn" and Readily Engage in The Exploration Process.
1. Children exhibit curiosity and ponder observations.
2. They move around selecting and using the materials they need.
3. They take the opportunity and the time to "try out" their own ideas.

Children Plan and Carry Out Investigations.
1. Children design a way to try out their ideas, not expecting to be told what to do.
2. They plan ways to verify, extend or discard ideas.
3. They carry out investigations by: handling materials, observing, measuring, and recording data.

Children Communication Using a Variety of Methods.
1. Children express ideas in a variety of ways: journals, reporting out, drawing, graphing, charting, etc.
2. They listen, speak and write about science with parents, teachers and peers.
3. They use the language of processes of science.
4. They communicate their level of understanding of concepts that they have developed to date.

Children Propose Explanations and Solutions and Build a Store of Concepts.
1. Children offer explanations from a "store" of previous knowledge. (Alternative Frameworks, Gut Dynamics).
2. They use investigations to satisfy their own questions.
3. They sort out information and decide what is important.
4. They are willing to revise explanations as they gain new knowledge.

Children Raise Questions.
1. Children ask questions (verbally or through actions.)
2. They use questions to lead them to investigations that generate further questions or ideas.
3. Children value and enjoy asking questions as an important part of science.

Children Use Observation.
1. Children observe, as opposed to just looking.
2. They see details, they detect sequences and events; they notice change, similarities and differences, etc.
3. They make connections to previously held ideas.

Children Critique Their Science Practices.
1. They use indicators to assess their own work.
2. They report their strengths and weaknesses.
3. They reflect with their peers.

01/25/91 VESP Practice Profile: The Student "it".
04/92 Revised

The Vermont Elementary Science Project is located at Trinity College, McAuley Hall, Burlington, VT 05401,(802)658-3664.
The VESP is a grant awarded to The NETWORK, Inc., Andover, MA, by the National Science Foundation.

Reprinted with permission from TEACHERS' LAB, November & December 1991.

Assessment: what is "IT?"

by Gregg Humphrey

"THE ANSWER is not to be found in books about tests and examinations, but in what is happening in classrooms every day." This observation by the well-known British researcher, Wynne Harlen, is central to our approach to assessment at the Vermont Elementary Science Project. It is essential that educators develop assessment strategies that are effective and that do not force teachers into unworkable modes of teaching.

The term "assessment" is applied to a wide range of methods by which information about students is gathered and appraised, including formal testing and analysis. Gathering information about children as a result of day-to-day, informal interactions is part of teaching and is equally a part of assessment.

For many teachers, administrators, and parents the word "assessment" brings to mind "testing." It is important to broaden and change this limited conception of assessment to a wider view which promotes the growth of science and math concepts, process skills, and attitudes that are inherent in effective science and math development.

A common misunderstanding is that assessment is something that requires additional and/or different activities — as an adjunct to teaching. To some it means that teachers have to stop teaching and withdraw from interaction with children in order to observe them. Yet with effective reading and writing instruction, teachers routinely make immediate assessments based on behaviors and abilities which students exhibit in the act of reading and writing. These judgments, or "continuous assessments," often result in probing questions and other pedagogical decisions on the part of the teacher. Such judgments are used in the service of instruction as well as to monitor and record both short and long term student development. Similarly, when students are doing science, teachers can learn to facilitate the activities as well as to assess conceptual, process, and attitudinal indicators congruently.

A fundamental piece of assessment occurs as teachers and students are engaged in the daily activities of hands-on, inquiry-based science. Just as teachers listen to students read and diagnostically raise questions and suggest strategies for improvement, teachers interact with children as they "do science" in order to determine their progress and plan for future instruction.

Assessment to benefit learning

We need to develop and adopt assessment techniques that match what is valued in science learning and that reinforce the goals of science education. For example, rather than just testing for factual recall, we should assess the extent to which students can apply the knowledge they possess. We must develop techniques to probe for depth of understanding of facts and concepts, and the ability to use problem solving skills.

A fundamental piece of assessment occurs as teachers and students are engaged in the daily activities of hands-on, inquiry-based science.

As classrooms increasingly use hands-on approaches for building students' understanding of scientific concepts, hands-on performance can be used to assess student competencies and skills. Further, we need to discard the view that assessment is only done after completion of a unit and for the purpose of giving students a grade. Assessment needs to be a means for improving instruction, not just measuring its impact. Day-to-day assessment can serve instruction by helping teachers monitor the progress of students in science so that adjustments to enhance student learning can take place.

GREGG HUMPHREY is the Technical Assistance Specialist for the Vermont Elementary Science Project. His article, "The Nature of Elementary Science: What Does "IT" Look Like?" appeared in the September 1991 issue of Connect.

"Assessment in education has been criticized for interfering with the process of learning, the analogy being that of the gardener constantly pulling up his plants to see if the roots are growing. There is some truth in this . . . but it also distorts reality to make a point. Gardeners do have to find out if their plants are growing and they do this, not by uprooting them, but by careful observation with a knowledgeable eye, so that they can give them water and food at the right times and avoid either undernourishment or over-watering . . . The gardener who does not know what size and shape a plant is to be and how quickly it is expected to grow will not know what signs to look for, and may mistake a condition which is quite normal for one which requires remedial action, or vice versa. We have to know something about the development taking place in order to interpret what we find when we assess it."
(p. vii)

Science: Guides to Assessment in Education, 1983, Wynne Harlen

BEST COPY AVAILABLE

Reprinted with permission from THE SCHOOL ADMINISTRATOR, December 1992.

What's Worth Assessing?

BY MONTE MOSES
Associate Superintendent, Cherry Creek Schools, Englewood, Colorado

Over the past several years I have been involved in alternative assessment projects as a district leader, principal, and consultant. These experiences have convinced me that high quality assessments can enhance student growth and achievement.

Along the way, I have discovered a few ideas that others might consider and use.

First, get clear on what alternative assessment really is. I advise colleagues to focus their energies on creating "performance assessments" as opposed to "alternative assessments" because the words indicate better the kind of measures that are needed in today's schools. Performance assessments focus on specific results, while alternative assessments could be virtually anything.

Design tasks that are worth assessing and that provide useful information on core district outcomes. Although it sounds easy, not enough time is spent defining the assessment task. The tendency is to look for things easy to measure instead of asking what is important to assess.

I encourage administrators and teachers to begin with this question, "What task could students perform that would be a strong predictor of later success in school and life?" These are the things we should teach, not just assess!

At Mark Twain Elementary School in Littleton, Colo., we reached agreement that the ability to research a topic of social or scientific importance is one such task. We viewed this task as very important because it calls for students to use skills such as reading and writing and to integrate knowledge from different subjects. Better yet, it is used repeatedly in all levels of schooling and walks of life.

We envisioned the research assessment as being much more in-depth than simply writing a report, which our students had done a great deal but with questionable quality. Our goal was to design an assessment task and scoring system that would get students to move beyond the plagiaristic tendencies so typical of elementary school children.

We discovered the "trick" to good assessment isn't getting students to do something completely foreign. Rather, it is a matter of being clear about what quality means.

Linking Instruction

The assessment tasks themselves must be engaging. Students must be interested in an assessment task before we can assume it is their best work, and then hold them accountable to high standards of thoughtfulness and self-direction. Thoughtfulness and self-direction demonstrate the student possesses more than just knowledge and skills. They show the student has developed the attitudinal disposition needed to apply them.

Administrators must help teachers develop a tight conceptual linkage among curriculum, instruction, and assessment. Without this linkage, teachers are likely to view performance assessment as an intrusion into instructional time and a diversion from "covering" the curriculum. The more productive mindset is to see performance assessments as "tests worth teaching to."

As performance assessments are developed, make sure everyone understands they exist to strengthen teacher judgment, not replace it. The information obtained from a performance assessment can be quite helpful, but it is still up to the teacher to interpret and use this information.

Another important step is to get students involved in self-assessment of their performance against predefined standards. Wouldn't it be wonderful if students could appraise their own work against standards in a manner identical to what the teacher would do? This idea is not at all far fetched.

Start now to create performance assessments. Personal experience is the key in this endeavor. Until teachers have struggled with the problems of task selection and setting standards, they will not fully appreciate the instructional value of good assessment. The more staff who are involved in these discussions, the better. Involving parents and community members is also a good idea. They can add a lot.

Don't be afraid to make revisions based on feedback from parents, teachers, and students. The research assessment at Mark Twain Elementary was revised substantially seven times. We had to swallow our pride, but the end result was worth it.

After working out the kinks on the research performance assessment, we institutionalized it. The assessment was given to students several times during their elementary school experience, in hopes they would improve from one try to the next. The assessment took on the character of "the big game" and motivated students to prepare well in advance. Better performance was the result.

Altering Perceptions

Among many obstacles, two stand

out. The first is resources. The development of performance assessments takes time, and time costs money. However, the initial funds required are small in comparison with what districts commonly spend on commercial testing, whose value often is questioned. Also, keep in mind that the investment in performance assessment is good staff development.

The second obstacle relates to adult perceptions of students. One perception can be expressed in the form of a question, "What will we do with the students who fail the performance assessment?" This question assumes some students cannot learn, grow, and achieve when expectations are high. School leaders must confront that dangerous assumption.

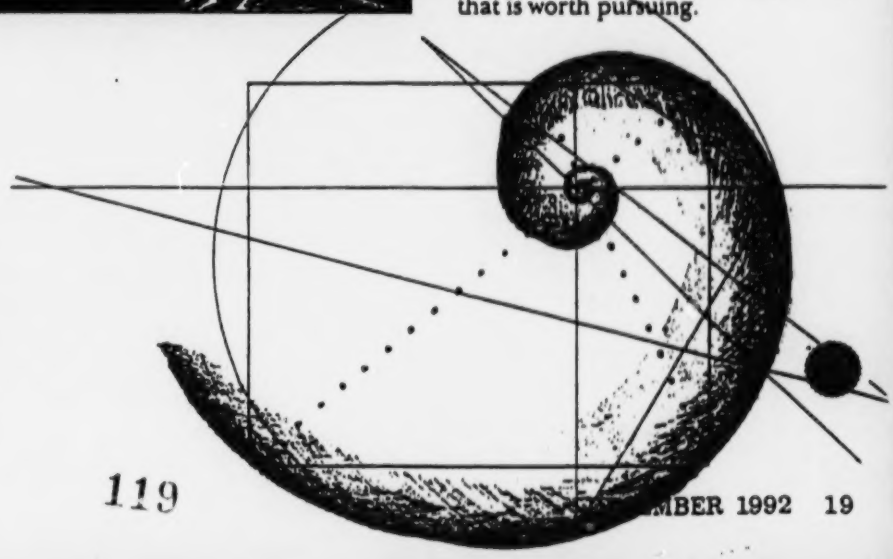
The question also fails to consider performance assessments are not one-shot tests. They are tasks worth repeating regularly because of their inherent value. Our job as educators is to coach students to higher performance from one try to the next.

Another obstacle concerns able students. The highest score on a performance assessment should be set beyond the current skill level of even the best students, an idea which will seem quite foreign to them. Many of these able students (and their parents) incorrectly assume that their work is excellent because it is better than their peers, when it may be far below standards of excellence. Strong students deserve a challenge that will call upon them to grow and stretch to meet high standards just as much as an average or below-average student.

Performance assessments give us a unique opportunity to raise standards and help more students to reach them than ever before. In my mind, that is worth pursuing.

Photos courtesy of New York State Education Department

Students participating in New York state's fourth grade science manipulative skills assessment are asked to measure objects using a ruler, double-pan balance, or a thermometer.



Reprinted with permission from Andrew Ahlgren, American Association for the Advancement of Science, Washington, DC.

Creating Benchmarks For Science Education

Andrew Ahlgren

Progression-of-understanding maps were one important tool in coordinating the research and information needed for creating K-12 science standards.

Project 2061 has been constructing goals for science, mathematics, and technology education since 1985. During our first three years of work, we recommended what students should remember by the time they leave high school (*Science for All Americans* 1989). Since 1988, we've been working on reasonable expectations for students at earlier grade levels (*Benchmarks for Science Literacy*, in draft). This new volume will include benchmark lists, some of our "progression-of-understanding maps," and essays related to the benchmark topics.

We intend the benchmarks to be used by school districts or curriculum developers in constructing alternative K-12 curriculum models adapted to their own populations and circumstances. Before we reach that point, though, we believe some reflection on how we created the benchmarks can be a stimulus to other curriculum reform efforts.

The experience of writing benchmarks is highly stimulating. But make no mistake about it: The work is difficult, and getting started seems to discourage many people from the undertaking. Still the quality of thinking and conversation that goes on is often impressive, even when the tentative product may not be.

Benchmark Grades

The National Assessment of Educational Progress (NAEP) has popularized grades 4 and 8 as benchmark grades, and the National Council of Teachers of Mathematics (NCTM) has followed that pattern (*Curriculum and Evaluation Standards for School Mathematics* 1989). However, the district teams working with Project 2061 decided that the psychological distance from K to 8 requires more than a single benchmark. The end of grades 2 and 5 were recommended as being more meaningful developmental breaks, and we are designing benchmarks for those grades.

It is not our intention that 2nd graders should be subjected to formal, national examinations on their progress in science (as seems to be in the cards for older children). The "feel" of the expectations for grade 2 is distinctly different from that for grade 5, so we believe it is important to discourage kindergarten teachers from embarking immediately toward grade 5 goals. By their modest, nontechnical nature, grade 2 benchmarks suggest what students cannot be counted on to know as they begin grade 3, and this may temper expectations for what children can learn in grades 3 through 5.

Inferring Benchmarks

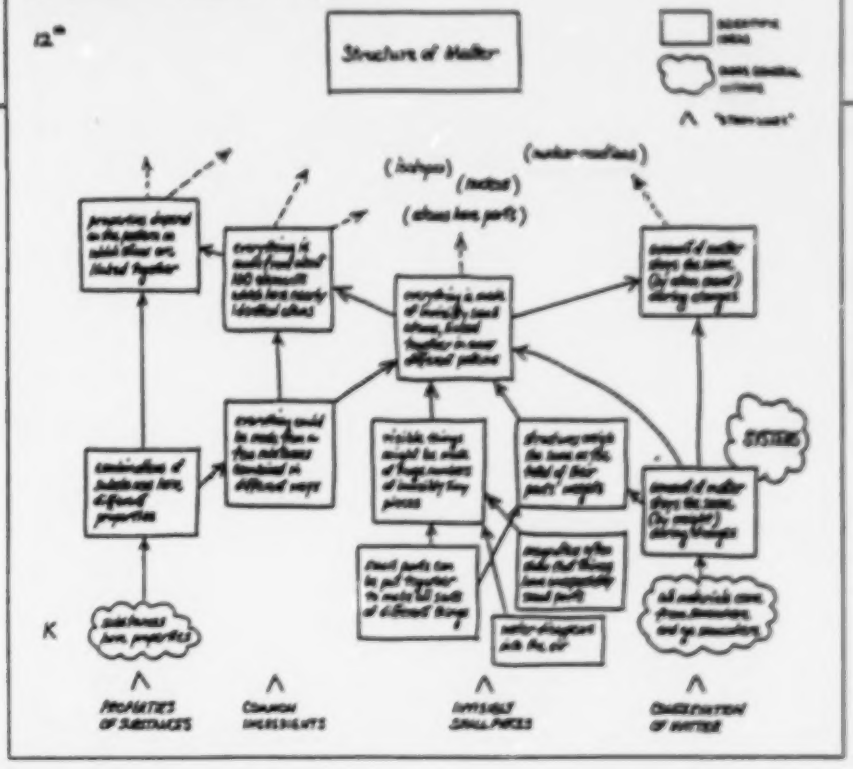
In crafting the lower-grade expectations, we drew partly on an analysis of what ideas would be needed to achieve the 12th-grade understandings in *Science for All Americans*. We also considered estimates of what students are capable of at different ages, drawing information from the experienced teachers on our district teams and from researchers who study how children understand and learn science. Unfortunately, the availability of published research on children's understanding of science is very uneven over different content areas.

We have found that it is seldom possible to work backward from 12th grade goals one at a time to create a neat stack of previous levels of sophistication. Usually there are convergences (several ideas required to understand a subsequent idea) and divergences (several ideas depending on one prior idea). The natural medium to express such goals is therefore a diagram with boxes and arrows. We called our speculative charting a "progression-of-understanding map" (to distinguish it from the "concept map" currently popular in science education).

Figure 1 is an example of a draft of a progression-of-understanding map, depicting ideas related to a section on the structure of matter (*Science for All Americans*, Chapter 4: "The Physical Setting"). Reading from bottom to top, the map shows a rough progression in time, beginning from notions students hold when they enter school. When we sketched this map, the *sequence* of ideas was more important than tying ideas to any particular grade. Estimation of approximate grade placement for each idea usually came later.

Figure 1

Progression-of-Understanding Map



Unfortunately, the availability of published research on children's understanding of science is very uneven over different content areas.

The toughest part of the map was in the middle. Without the firm guidance of research on many topics, we were in the same place as the Geography Task Force of the National Council on Education Standards and Testing when they wrote: "It was difficult to set 8th grade standards, other than indicating that students should be expected to know more than they did in the 4th grade and less than the 12th grade" (*Raising Standards for American Education* 1992, p. L-4). Compounding our uncertainty was the possible difference between what children could be expected to do now, with their current history in the school system, and what 5th or 8th graders eventually might be able to do if they had optimal experiences.

When we began mapping, we intended to cover one conceptual strand at a time, leaving some loose ends that would later connect to other strands. For example, the structure of matter shows obvious connections to the flow of matter and energy. The structure of matter was a whole section in *Science for All Americans*, and perhaps it was too large a conceptual chunk to represent comfortably on a single map. (Notice that it is incomplete in fig. 1.) It soon became evident that a progression-of-understanding map for a single strand was already more complex than most people would find inviting.

Software support for constructing each map and for making connections among them would be very helpful (and we now have a grant to develop such software). We intend to create a curriculum resource data base that will link appropriate parts of the progression-of-understanding maps, giving users the option of choosing the complexity of information they want to consider. The resource base would also link benchmarks to blocks, to activities and materials, and to appropriate assessment suggestions.

We plan to accompany benchmark lists with essays that will call attention to the progressions of various strands and connections among them. For example, the essay accompanying the benchmarks for the structure of matter would draw attention to the parallel development of four different strands: properties of substances, combinations of parts, invisibly small pieces, and conservation of matter.

Essays, which are being prepared by our teams and staff, will also draw on the available educational and psychological research, calling attention to difficulties that students are likely to have at each level and, in particular, to persistent previous

Figure 2

Literacy Goal: The Structure of Matter

The following example of an essay and benchmark list is taken from the draft of Benchmarks for Science Literacy *(in draft).*

Students will learn about the nature of atoms and molecules and the structure of matter.

Of all sections, this one may have the most implications for students' eventual understanding of the picture that science paints of how the world works. However, it may also offer the most difficulties. The theory of atoms and molecules is powerful in explaining our world, but it requires bringing together a number of lines of evidence and imagination: about the properties of materials and their combinations, changes of state, effects of temperature, behavior of large collections of pieces, the construction of objects from parts—even about the desirability of simplicity in explanation. All of these should be grasped by children during middle school, so that the unifying ideas of atoms can be developed by the end of grade 8.

The scientific understanding of atoms and molecules requires students to entertain the notion that all visible things are composed of invisible particles. Another notion is that everything might be made up of a relatively few ingredients. An idea preliminary to this is that materials combined in different ways can have different properties. And still preliminary to that is the very notion of "properties" of materials.

Parallel to consideration of properties of combinations is the notion that the bulk properties of materials can be very different from the properties of their minute parts—an idea counter to the students' intuition.[1] And parallel too is the idea of an unchanging total amount of matter, beginning with the evidence that total weight stays the same during all sorts of changes in materials.[2]

Grades 3 through 5

The study of materials should continue throughout these years but become more systematic and quantitative. Students should design and build things that put different requirements on the properties of materials. They should be expected to write clear descriptions of their designs and experiments, present their findings whenever possible in tables and graphs (designed by the students, not the teacher) and enter their data and results in a computer database.

Students should measure (weight, dimensions, temperature), estimate (dimensions, weight, population size), and calculate (area, volume, population size) using hand-held calculators when necessary.[3] With magnifiers, they should inspect substances composed of large collections of particles—sand, spices, powders—to discover the unexpected details at smaller scales. They should observe and describe the (sometimes solid-like, sometimes liquid-like) behavior of large populations of pieces—powders, marbles, sugar cubes, or wooden blocks.

By the end of the 5th grade, students should know that:

- Heating and cooling cause changes in the properties of materials. Many kinds of changes occur faster under hotter conditions.
- However parts are assembled, the weight of the thing made is always the same as the sum of the parts; and when a thing is broken into parts, the parts together weighed the same as the original thing.
- Materials may be composed of parts that are too small to be seen without magnification.
- When a new material is made by combining two or more other materials, it can have properties that are different from any of them. For that reason, a lot of different kinds of materials can be made from a small number of basic kinds.
- A collection of a large number of pieces may keep its shape or flow like a liquid, depending on how the pieces stick together or how they are stacked.

[1] Brook et al. 1984, Driver 1987.

[2] At the beginning, children have various ideas about what is "matter." For many, gases and even liquids are not seen as having weight or as being matter (Lee et al. in press, Driver 1987, Stavy 1990). Very tiny solid particles are also not seen as having weight—because their weight cannot be felt (Smith, Carey, and Wiser 1985, Carey 1991)..

[3] Research shows that children may consider anything so light that they cannot feel its weight to have no weight at all (Smith, Carey, and Wiser 1985; Carey 1991). Lots of weighing on increasingly sensitive balances, including weighing piles of small things and dividing them to find the weight of each, will help.

conceptions that may interfere with learning (see fig. 2). We are still uncertain about how far essays should go beyond suggesting appropriate kinds of instruction. (The research has much less to say about instruction to overcome difficulties than about the difficulties themselves.)

Benchmark Adjustments

Once in the thick of producing benchmarks, occasions will arise when a benchmark statement doesn't seem well suited to its designated grade level. The easiest option is to move the benchmark intact to another grade level, making adjustments to any benchmarks connected to it.

A second option is to rewrite the statement at a level of sophistication more appropriate to the current grade level, but this is more than a stylistic transformation. Grade adjustments can seldom be made so simply as changing the vocabulary. Rewriting usually requires rethinking what students should be able to do—in their heads or behaviorally—at that level.

A third rewriting option is the most difficult but probably the most fruitful: tease apart the substance of the benchmark and create two new ones, keeping one at the current level and putting the other at a different one. Again, merely making style changes in language won't change the substance of the benchmark. One must reconsider what students could understand and what the likely sequence of understanding is.

Knowledge vs. Belief

Research shows that children may *understand* a scientific explanation of phenomena before they *believe* it (for example, Hewson and Hewson 1992, Osborne and Freyberg 1985). The longer the time gap between being able to state an idea and eventually believing it, the greater the problem for writing benchmarks. Should a benchmark about children's ability to explain something specify that they can produce a scientific explanation, or should we try to require their acceptance of it as well?

From a philosophical point of view, Project 2061 would prefer to require knowledge rather than belief. A similar dilemma appeared in writing the "Values and Attitudes" section in *Science for All Americans*. We rejected the goal that everyone should like science, mathematics, and technology or should believe these endeavors are of net benefit to humankind. We agreed instead on the goal that students' attitudes—whether they turn out to be positive, negative, or neutral—should be based on a sound understanding.

A poignant case might be that of evolution through natural selection. We can reasonably require that students understand what the scientific theory is, but do not have to require students to believe that is how present life on earth necessarily came to be.

Final Thoughts

Benchmark drafts should be tried out with a variety of readers, not just for approval or minor editing, but to see how they are likely to be interpreted and used. Writing good benchmarks may not require setting fixed rules as much as it requires being continually vigilant about how one's intent might be misunderstood.

Researcher Pat Heller (of the University of Minnesota) summarized the task for us after a recent writing retreat:

- Make benchmarks not so specific as to be limiting and not so general that no one is quite sure what you're talking about.
- Have a clear sequence where necessary within a grade level.
- Have a progression from one grade level to the next that illustrates increasing sophistication.
- Show connections between benchmarks under different goals.
- Write them to be developmentally appropriate, assessable, and relevant to the child's world. ■

References

Hewson, P., and M. Hewson. (1992). "The Status of Students'Conceptions." In *Research in Physics Learning: Theoretical Issues and Empirical Studies*, edited by R. Duit, F. Goldberg, and H. Niedderer. Kiel, Germany: Institute for Science Education.

Osborne, R., and P. Freyberg. (1985). "Roles for the ScienceTeacher." In *Learning in Science*, edited by R. Osborne and P. Freyberg. Auckland: Heinemann.

Raising Standards for American Education. (January 24, 1992). Washington, D.C.: National Council on Education Standards and Testing.

Science for all Americans. (1989). Washington, D.C.: American Association for the Advancement of Science.

Copyright © 1993 by Project 2061.

Andrew Ahlgren is Associate Director, Project 2061, American Association for the Advancement of Science, 1333 H St., N.W., Washington, DC 20005.

Reprinted with permission from NSTA Publications, copyright March 1991, Science and Children.

Assessment, Practically Speaking

How can we measure hands-on science skills?

By Lehman W. Barnes and Marianne B. Barnes

THOUSANDS OF ELEMENTARY school children busily do hands-on science activities every day. They observe and classify objects, measure length, volume, and mass, collect and record data, and manipulate materials. These students know that science involves both doing and communicating. But as they prepare for science tests, what have students come to expect? Are they being assessed on *all* that they've learned?

The Motor-and-Wires Kid

Jimmy was thrilled when we began a science unit on electricity and magnetism. I had assigned the first 11 pages of chapter four and was beginning to discuss the main ideas of the unit. Suddenly Jimmy burst out, "I have a couple of motors at home, could I bring them into class and show them to everyone? Or one of the 'electric magnets' I made last summer? I saw a picture of one in a book about science at the grocery store. All you need are some wires, a nail, and a battery."

"Sure Jimmy," I said, glad to see him excited. "Bring in your motors and wires."

The next day, as I was unlocking the class door, I heard the rush of footsteps. There was Jimmy with a large paper bag containing all kinds of stuff—batteries, motors, wires, nails, pliers, and tape. "When are we having science today?" he asked me. His excitement was evident as he pored through his bag.

DANE PENLAND, SMITHSONIAN INSTITUTION

Doing a hands-on biology activity.

Teachers thrive on flexibility, especially in response to student enthusiasm, so I began science right after the morning announcements, a time usually reserved for language arts. Jimmy showed his classmates the motors, batteries, and wires, demonstrated how to build an electric magnet, and built an electric circuit. He eagerly shared the contents of his bag with the other students, most of whom were equally eager to learn what he knew about the motors.

During the next two days, with the class's interest and excitement levels staying high, I introduced the main ideas of the science unit and, thanks to Jimmy, got the whole class involved in creating series and parallel circuits. I ended the activities with a brief review for Friday's science test.

What Does My Test Test?

The comprehensive test—vocabulary, labeling, matching, multiple-choice, true-or-false, short-answer, and puzzle questions—accurately assesses students' mastery of the verbal aspects of the science unit. Jimmy did well that Friday, especially on the short-answer questions. His answers were complete, and he even included drawings of his two setups. At the bottom of his test paper he wrote, "I think I know a whole lot more, but I can't show you on this kind of test."

Jimmy's comment struck me. The test lacked an opportunity for students to demonstrate what they had learned

through the manipulation of materials. How do test items assess practical performance? Paper-and-pencil tests adequately assess verbal learning, but they often neglect the equally important practical learning.

The quantity and quality of practical learning experiences in elementary classrooms has grown in recent years. Practical assessment, however, lags behind. If we believe in hands-on learning, then we must approach the assessment of this learning with practical hands-on testing. Otherwise, we convey the message that practical learning is less important than verbal learning.

Evaluation Potential

Practical assessment tests hold an enormous potential for elementary science teachers, because so many areas of science would be better evaluated in such a manner. For example, think about evaluating science skills in the following scenarios:

- working with basic science equipment, such as a thermometer, a triple beam balance, a meterstick, a graduated cylinder, and a stopwatch;
- performing laboratory skills and procedures, such as working with a magnifying glass or microscope, heating or filtering a substance, mixing a solution, or measuring rates;
- observing and classifying three-dimensional objects, such as shells, rocks, plants, and animals (currently, most assessment tests employ visual observation only in two-dimensional settings);
- collecting and recording data in tables, charts, and graphs that students create themselves—such graphs could reflect temperature changes over time, heartbeats per minute, or a ratio of the manipulated variable to a responding variable;
- designing experiments and performing investigations from a set of materials and a specific question (Which paper towel will hold the most water? How can you make the tablet dissolve faster? Which material would keep you warmer?);
- generating a set of investigative questions from a set of materials, such as a cart, weights, and an inclined plane; a set of paper gliders; or a set of pendulums;
- manipulating objects to demonstrate understanding of concepts and connections among concepts—for example, understanding electrical circuits, or understanding the relationships between the mass, volume, and density of an object;
- building models—physical representations that demonstrate natural phenomena—such as a cell, the solar system, or a geological structure;
- communicating in the process of investigation during small group work, individual writing, summarizing and presenting to others, or creating a display.

IRENE FENLAND, SMITHSONIAN INSTITUTION

Science assessment should measure hands-on skills.

Verbal *and* practical learning are fundamental ingredients of a good elementary science program. Historically, verbal learning has been emphasized in schools and has, appropriately, been measured by verbal assessment. Now, with practical learning finding its rightful place in the instructional program, practical assessment should develop as well. Elementary educators must face the challenge to practically assess their increasingly creative hands-on science teaching.

LEHMAN W. BARNES is an associate professor of science education at Florida State University in Tallahassee. MARIANNE B. BARNES is a professor of science education at the University of North Florida in Jacksonville. They are coeditors of the Florida Science Teacher.

BEST COPY AVAILABLE

Reprinted with permission from HANDS ON!, Spring 1991, published by TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140. If you would like actual copies of this issue, call 617-547-0430.

Getting Connected to Science

by Candace L. Julyan

Consider how you first became interested in science. Perhaps your interest grew from a curiosity about a particular phenomenon or organism. Satisfying that curiosity, or what David Hawkins (1967) calls "messing about," often resulted in some type of relatively unstructured exploration led by your own questions rather than the questions of others. Unfortunately, messing about is not a common practice in many science classrooms today. Students are more likely to be introduced to organisms and phenomena through text-based lectures rather than through their own observations. The result of our current science curricula is not only a lack of interest in science as a profession but also a lack of scientific understanding.

Data from numerous studies (such as Bell and Brook, 1984; Duckworth et al., 1985; Julyan, 1988) suggest that knowing correct terms does not help students make sense of their observations. In fact, in many cases scientific terms may be used to mask confusion.

Although I am certainly not suggesting a classroom ban on scientific terms, I do suggest that we reconsider their importance. The difference between memorizing vocabulary words related to science topics and understanding various phenomena is considerable. Certainly those of us involved in science and science education realize the difference and are striving for the latter. Science-as-vocabulary requires less effort on the part of both the teacher and student, but it also provides fewer rewards.

Students are also aware of the difference and, if given the choice, they too would strive for understanding. This point was highlighted for me several years ago by a high school student who was participating in a research study on how students make sense of a complex system (Duckworth et al., 1985). To examine this research question, my colleagues and I had devised a number of experiments featuring helium balloons weighted with enough strands of string so that the string dragged on the floor. The students' task was to explain why a balloon's position in the air changed as they did various tasks such as tying knots in the string, cutting the string, and putting the balloon on a table.

Early in the study, one student made it clear that she would not accept a scientific term as an explanation for what she was seeing. Several students consistently used the word "gravity" to explain anything that they could not easily make sense of otherwise. This particular student protested that explanation vehemently, exclaiming that she had never understood what gravity was and that the word certainly did not help her now as she worked with the balloons.

After several weeks of experimenting, the balloons continued to surprise her, and her frustration at her lack of understanding was often visible. One day, she turned to me demanding to know if I intended to tell her "the answer." When I asked her to state the question, she seemed momentarily stumped but then stated that she wanted to know why the balloon behaved in the way that it did in all the various experiments she had conducted. Although certainly willing to answer her question, I said that I wanted to be clear about what she wanted. Did she want words that explained the phenomenon or did she want to understand it herself? There was a long and poignant silence in the room. Finally, she turned and quietly said that the words were *not* what she wanted, she really wanted to understand.

Although most students might not articulate the dilemma as clearly as this student, many share her desire to understand, not just to know, the correct words. One way to promote both understanding and the value of scientific terms is to give students an opportunity to mess about in their science classes, to become participants in constructing their knowledge. Inquiry-based activities are certainly not the fastest way to approach science learning. The faster and more efficient approach is definitely text-based, lecture-based

Dr. Candace L. Julyan is Project Director of the NGS Kids Network and Principal Investigator/ Project Director of NGS Kids Network/Middle Grades.

classes. However, this more common approach does not appear to be terribly effective. Students' notions about scientific topics often differ from those of scientists and form a coherent fabric of understanding that has its own internal logic. A presentation of new, sometimes conflicting, information by a teacher does not necessarily change, or even challenge, students' initial ideas about the topic. This is an important point to keep in mind — students do not learn simply by being told.

One of TERC's goals through the years has been to encourage educators to consider the possibilities and the strengths of "messing about" in science and mathematics. For the past several years we have taken a new perspective on student inquiry through the projects we call "network science." The basic premise of these curricula is that students can and should *be* scientists, that they can and should converse with real scientists about their work, and that computers can enhance this enterprise. Students conduct experiments, analyze data, and share results with their colleagues using a simple computer-based telecommunications network. This collecting and making sense of data gives the students an opportunity to experience the excitement of science that scientists feel.

Our network projects are diverse, from National Geographic Kids Network to Star Schools to Global Lab. Although each program has unique characteristics, as a group they represent an exploration into the power of a particular way of messing about. Most of the units in these supplemental curricula focus on an environmentally-oriented topic, such as acid rain, trash, water quality, or radon. Students usually begin their study by examining the topic in the context of their local community and then expand their inquiry by exploring it within the larger national and international picture. Students collect some type of data (survey or measurement), share those data through telecommunications with other classes collecting the same data, and finally make sense of those data using computer (and non-computer) tools. As much as possible, we have a scientist, a professional with expertise in the unit topic, available to the students to help them consider the questions raised by the data.

Photo by James Gee

TERC Star Schools students working intently on their group project.

We view the computer as an important tool in these inquiries, providing students with help in both data collection and analysis. The typical computer configuration features a word processor, a record-keeping data section, a graphing utility, a complete telecommunications package, and map software with data overlay.

Our network science projects have generated an enthusiastic response on the part of both teachers and students. One NGS Kids Network teacher explained network science as "an awesome concept, a truly revolutionary idea for education at a time when it is so badly needed." At present we have tested these materials in approximately 5,000 classrooms across the United States and at a growing number of foreign sites, including Argentina, Australia, Canada, Hong Kong, Israel, the Soviet Union, and Zimbabwe. Data from our various evaluations suggest that network science is a lively and appealing way to approach science. From these data we have come to have a number of "beliefs" about the possibilities of this approach. We offer them as points for consideration and discussion for a wide audience of educators, from teachers to administrators to other developers.

Data collection and analysis provide an effective backbone around which to study science.

In addition to the expected learning about the particular subject covered in the units, teachers and students reported numerous examples of scientific thinking generated by the simple act of collecting and making sense of data. In many instances, units begin by asking students to consider definitions, not as vocabulary exercises but as an important aspect of understanding data. Students in one fifth-grade class, preparing to collect data on trash, spent considerable time deciding whether "used" was different from "you don't need it anymore." The longer students discussed the difference, the more they realized the connection between trash and people's personal habits. This class finally agreed that trash was "something that has used its purpose."

This definition matched the thinking of some other classes that defined trash as "something no longer useful to a person" or "useless, worthless, worn out, unwanted, and of no value to you." Other classes' discussions about trash seemed to look at trash in a slightly different way as they considered trash in relation to waste reduction solutions.

These students defined trash as "any object that a person doesn't want or has no further use for, or that is not being
ycled, reclaimed, repaired, or
graded or that may pollute the earth" or "waste, scraps, garbage, litter, old things that are no longer in use, spoiled and ruined things, and things that can be recycled but weren't." As students continued the unit and shared data about the actual trash collected, they also considered the ways in which students' and communities' ideas about trash might affect the actual data. The exercise of considering all the complications of data collection and analysis is a vital step in the students' understanding of science.

Another example of the type of scientific understanding that data collection generated came from a fourth-grade class. These students realized that although their data and those of their teammate school were both about pets, they could not com-

Photo by George Miselis

A hands-on activity of the NGS Kids Network.

pare data. One class had collected the number of students who owned each
1 of pet; the other had collected the
nber of each kind of pet. The class learned a great deal about the importance of comparable data, a concept that many high school students might find difficult.

Technology greatly enhances classroom inquiries.

In many ways technology can provide an important link between science in practice and science in classrooms. Through the power of telecommunications, students are motivated to collaborate and share data and ideas with other students from all around the world. By using microcomputers for computations, students are able to manipulate their findings and explore their ideas in more sophisticated ways than their computational skills might have permitted otherwise. And telecommunications offers a unique and manageable opportunity for scientists to communicate with science classrooms. The technology expands these classrooms by eliminating the limitations of both time and distance that would otherwise restrict this type of communication.

Students involved in a network science project about radon were able to collect readings in their town, compare those readings with other towns around the country, and converse with a scientist at the Environmental Protection Agency about their questions regarding the differences they found in wooden and stone houses. A text-based study of radon would not have provided them with the variety of opportunities to consider, reconsider, and talk about their ideas on this topic. In this case, technology provided the tools for analysis, as well as an opportunity for expanded communication.

Network science helps a larger proportion of students feel confident about their ability to understand science.

This point is addressed in several ways. Many teachers reported that students were motivated and eager to participate in the curriculum activities, even, or perhaps especially, those students who rarely participated in science class. One teacher told the story of a "low ability" student who gained enormous credibility in class when he proposed his idea about the wide discrepancy between the over 120 pets owned by his Auburn, Maine, class and the fewer than 25 pets owned by a class on his research team in New Orleans. Whereas other students had explained the difference as something related to the weather or the availability of pets in Louisiana stores, this student thought that perhaps the school was located in an area surrounded by government housing. He explained that pets are not allowed in this type of housing. This simple piece of information from a fellow student, not from the teacher or a textbook, helped students make sense of the data and generated a letter to find out if this student's theory was correct. As you might imagine, his idea also increased the student's credibility with his peers, his teacher, and perhaps himself.

We received other reports of increased student self-esteem that are of a very different nature. Teachers reported that students felt a real sense of ownership about their work. The most extreme examples came from two teachers who reported that their classes, both filled with "average" and "below average" students, protested upon learning that their teacher was planning to present the NGS Kids Network curriculum at a professional conference. These students argued that because it was *their* work, they should be the ones presenting. In one case the teacher took the students with him; in the other, the students made an acetate filmstrip and accompanying audio tape that the teacher used in her presentation.

Science teachers participating in our network science programs consistently report increased involvement in classroom work by all their students, particularly those who previously had been academic underachievers. One teacher in the Star Schools project reported that one such student "has taken a complete change in personality and performance after his [solar] house was one of the most successful. He just displays a confidence that I didn't see before this unit. Learning styles were so evident to me while observing my students complete the unit. It gives me

new insight into better teaching methods."

Students' work in classroom inquiries is of interest to the local community.

Network science classrooms are filled with visitors interested in the students' work. The list is long and varied and includes principals, other teachers, superintendents, school board members, parents, university professors and their students, and both television and print journalists. These visitors are interested in what the students are doing, both in terms of the telecommunications activities with other schools and the data they are collecting, particularly when the data are about larger environmental concerns such as acid rain. This type of interest motivates both teachers and students, who appreciate having an audience for their work.

The work of a science classroom is of interest to the scientific community.

The scientific community has been both supportive of and enthusiastic about all our network science projects. Many scientists have given hours of their time exploring the types of research students might do that would offer valuable data to existing studies. Dr. John Miller, the deputy director of the acid rain disposition unit of the National Oceanic and Atmospheric Administration (NOAA) and the unit scientist for the NGS Kids Network Acid Rain unit, is an excellent example. (See "Scientists in Science Education: An old idea with a new approach" in *Hands On!*, Spring 1989.) Dr. Miller corresponded throughout the unit with his elementary-level "colleagues" and included the NGS Kids Network data as an appendix to the NOAA 1988 report on acid rain. In addition, he regularly discussed this project with his colleagues around the world.

Perhaps the most compelling example of the serious interest that scientists have in student-collected data took place at the annual meeting of a United Nations steering committee concerned with the long-range transmission of air pollutants. As the United States representative to this group, Dr. Miller presented an overview of the various networks in North America that are concerned with acid rain. Within this context, he introduced his colleagues to the NGS Kids Network project and was surprised to find that they expressed considerable interest in the network, wanting to know how students made the measurements and if the data would be available for comparison with NOAA data. We have found that this level of professional interest is not unusual. Scientists are interested in student measurements, particularly when the data cover a wide geographical area.

In these times filled with reports of the grim realities of science education, our work with network science projects has given us considerable hope about the possibilities for real change. Instructional materials that center around the collection and analysis of data not only give students a more accurate picture of the scientific enterprise, but they also give students an opportunity to examine and construct understanding of the topic under study. This approach offers them a collective opportunity for messing about in science. In addition, if this type of investigation is linked to real environmental concerns, the students' work becomes valuable community information, not just an empty school assignment.

In these times filled with reports of the grim realities of science education, our work with network science projects has given us considerable hope about the possibilities for real change.

One high school student involved in a Star Schools project described what he felt was the difference between his usual science class, which was lecture- and text-based, and his work in a network science study. "It wasn't like writing data charts, because data charts are really easy, you just copy things from the board or you do math. But [in] this [program] you have to sit down and think. You had to use your brain a little more than usual, so that's how it was different.... I think if we hadn't done the radon activity, I wouldn't have to use my brain at all the whole year. This is the only time that I really had to think.... I had to struggle a little, you know, and I'm not really used to doing that in science."

References

Bell, B.F. and Brook, A. (1984). *Aspects of secondary students' understanding of plant nutrition.* Leeds, England: Center for Studies in Science and Mathematics Education, University of Leeds.

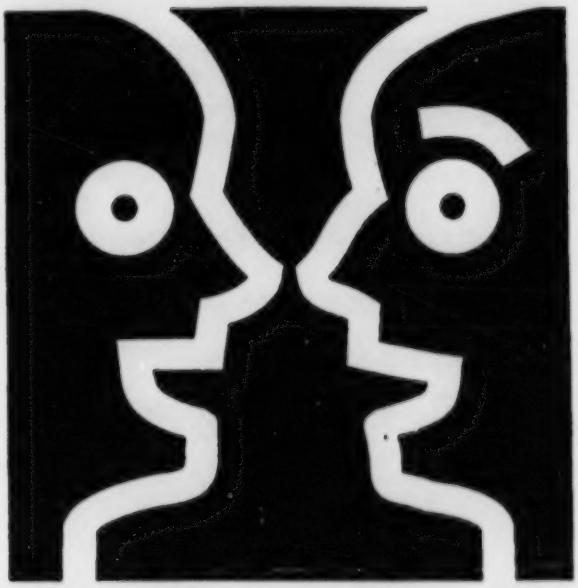
Duckworth, E., Julyan, C., and Rowe, T. (1985). *A study in equilibrium: A final report.* Cambridge, MA: Educational Technology Center.

Hawkins, D. (1967). Critical barriers to science learning. *Outlook*, 29.

Julyan, C. (1988). Creation of a Curriculum. *Hands On!* 11 (1), 6, 20.

Reprinted with permission from Science Edtalk, Council for Educational Development and Research, Washington, DC.

EDTALK



What We Know About Science Teaching And Learning

What role does assessment play in effective science instruction?

Assessing student progress and instructional quality is an important part of science education. One thing must be made clear from the outset: assessment encompasses more than testing, and much more than standardized testing. It includes such techniques as systematic teacher observation and so-called "authentic" assessment, in which the tasks assessed more closely parallel the learning activities and outcomes that are desirable in the science classroom.

Assessment fulfills many functions, from helping teachers make day-to-day classroom decisions to providing parents, the public, and policymakers with information about the overall effectiveness of a science program. In the classroom, most teachers use combinations of teacher-made or end-of-chapter tests and their own professional judgments for day-to-day instructional decisions. They use standardized tests for broader assessment purposes.

New approaches to science teaching and learning have expanded the role of assessment. Consequently, some teachers, schools, districts, and states are trying out a broader range of assessments and using assessment in new ways. For example, in an effective science classroom, a teacher might use assessment to fulfill any or all of the following purposes:

New approaches to science teaching and learning have expanded the role of assessment.

- Find out what students know and can do before instruction begins.
- Determine how they are progressing toward learning goals.
- Identify which strategies and thinking processes students use to reach answers or conclusions.
- Pinpoint which questions to ask to determine how well students are integrating new information.
- Establish what students have learned after a specific period of instruction.
- Motivate students.
- Inform parents about their children's learning.
- Evaluate the effectiveness of special interventions.

- Signal to students the areas in which they need to improve.
- Communicate teacher expectations about what is important.
- Determine whether they need to alter their teaching.

Schools, school districts, states, and the federal government also use assessment to monitor the effectiveness of teachers, schools, districts, and special programs, and to make policy decisions.

When used simplistically or inappropriately, assessment can... negatively affect students' perceptions of themselves as science learners.

Assessment can be a powerful influence on curriculum and instruction, for good and for bad. The format of the assessment and the uses to which results are put can guide how teachers teach and students study, especially when applied to "high stakes" decisions such as allocation of resources, admission to special programs, or receipt of a high school diploma. When used simplistically or inappropriately, assessment can drive teaching and learning in unhealthy directions and can negatively affect students' perceptions of themselves as science learners.

Many researchers have become deeply concerned that assessment is not being used well in most science education programs. Concerns center around whether assessment instruments, such as norm-referenced, standardized tests, are being used for too many purposes for which they were not designed, and whether the results of tests are being misunderstood and misapplied. Some researchers have asserted that the most common assessment formats, particularly conventional standardized tests, reinforce outmoded or ineffective instructional practices. For these and other reasons, many argue that assessment is an area of science education that is ripe for reform.

How do standardized tests affect science teaching and learning?

Paper-and-pencil tests, typically standardized multiple-choice tests, are the primary assessment tools in most science classrooms. Standardized tests have advantages of efficiency for testing large numbers of students, and many people perceive them as a more objective and reliable measure of achievement than assessments that rely more heavily on individual human judgment.

Recently, however, standardized science tests have come under strong criticism. Many researchers argue that conventional multiple-choice tests are questionable proxies of science knowledge. Because they do not ask students to generate their own answers, the tests do not measure scientific thinking. And because they have a single right or wrong answer, they reinforce the misleading conception of science as a static body of facts. Further, they do not assess laboratory skills at all.

Many of the standardized science tests are norm-referenced, ranking individual performance against that of a larger group. This, too, concerns some analysts, who say it perpetuates the notion that only a few students — the top scorers — are smart enough to pursue science. Researchers also raise questions about how well standardized tests measure the performance of students who are outside the cultural mainstream.

Of even greater concern to many researchers are the ways in which standardized tests may drive instruction in less effective directions or affect children in negative ways. Some researchers contend that conventional standardized tests, by sampling a breadth of content in a superficial and unconnected way, actually reward instruction that drills students on low-level facts and name recognition. In a high stakes testing situation, some studies have found that it is not uncommon for teachers to give teacher-made tests that imitate the format of large-scale assessments. Even schools that have implemented many of the reforms described in this document find their efforts undermined when standardized test scores are what "count."

Because they do not ask students to generate their own answers, the tests do not measure scientific thinking.

Responding to these criticisms, researchers and test publishers are developing new forms of tests, including multiple choice tests that require students to justify their answers or that assess thinking processes, open-ended answer tests, and performance assessments.

What are some good or promising methods of assessment being used in science education?

Aware of the limitations of conventional tests, researchers and practitioners are developing and implementing new forms of assessment that seek to better reflect the goals and processes recommended for effective science instruction, although these new assessments have their own weaknesses.

Several of these new forms fall under the umbrella called performance assessment, in which students create an answer or product that demonstrates their knowledge or skills. Many of these are hands-on in nature. Students might be asked to do a laboratory experiment or to solve a real-life problem. Through a series of systematized observations and questions, teachers or outside evaluators would observe and evaluate both the processes the students use and their understanding of the major concepts involved. In this type of assessment, students may work together or separately, using the equipment, materials, and procedures they would use in good, hands-on science instruction.

Students might be asked to do a laboratory experiment or to solve a real-life problem.

Another type of performance assessment that works well for students who are verbally or visually oriented is the presentation, which could be in the form of an oral presentation or a poster paper. This latter format mirrors the way many scientists present findings. Authors create posters explaining ideas; these posters might include graphs, photos, maps, or drawings, as well as some text. For example, students might develop a poster that explains how a local river evolved to its present state and how it might be protected.

A related form of assessment is systematic teacher observation, in which teachers scientifically observe and record student behaviors as they carry out meaningful tasks. Still another type of performance assessment being pioneered in science requires students to amass a portfolio of their work over time, which might include lab notes, science journals, and other written products, or graphs and charts of laboratory findings.

Research literature is also giving greater emphasis to student self-assessment and peer assessment techniques, in which students — for themselves or others — reflect on experience, attempt to understand what took place, and make judgments about what was learned.

Integrated computerized systems, which track and report student performance at the same time students are learning, is a growing approach to assessment. Paper-and-pencil tests are also being revamped to allow for more open-ended and student-constructed responses.

All of these assessments attempt to meet criticisms of current testing practices and to focus attention more on processes and thinking skills. Many of the performance assessment models meet the added criteria, emphasized in research literature, that assessment should closely resemble instruction and differ only in purpose. In addition, they provide much richer knowledge on which teachers can base instructional decisions.

Performance assessments have drawbacks, however. First, they require staff development for teachers who will implement, score, and use the results of the assessments. Second, they are considerably more expensive to administer than conventional tests. Third, they take more time from the school schedule to implement — some may last over several days. Fourth, although it is possible to standardize performance assessments to allow for comparisons among groups, it is a demanding process and is still in the experimental stages for many types of assessments.

Finally, it is difficult to correlate performance assessments from one task to another or to use performance assessment to project future student performance. And that is one major function of standardized testing.

Many of the performance assessment models meet the added criteria...that assessment should closely resemble instruction and differ only in purpose.

How can science assessments be better linked to instruction?

Assessment that does a good job informing instruction should have several features.

Some researchers and educators look forward to the day when there will be a "seamless web" of instruction and assessment in which assessment is no longer be a distinct activity but is "built into" students' regular learning experiences, virtually indistinguishable from instruction. In this vision, students and teachers receive ongoing feedback as needed during the instructional process. Assessment for classroom purposes (as opposed to external monitoring purposes) is a routine, non-threatening process.

This vision calls upon students, teachers, and administrators to look at assessment in a different way. Teachers must understand that the primary information will not be quantitative, but will provide a rich portrait of student strengths and weaknesses. In addition, assessment to inform instruction does not have to compare students to one another, which means that student approaches and responses to a problem may look very different and still be correct. Students, especially high school students, will be encouraged to be the primary users of assessment information in their own learning.

Assessment that does a good job informing instruction should have several features. It should measure the processes students use as well as the answers they reach. It should measure all of the goals of the curriculum, not just a few. It should address both group activities and individual ones. It should be developed by, or with ample input from, teachers and should include teacher professional observation and judgment. Perhaps most importantly, it should draw upon information from multiple assessment sources, including but certainly not limited to tests.

Finally, assessment to inform teaching and learning should have a strong self-evaluation component for both students and teachers.

Reprinted with permission from EDUCATIONAL LEADERSHIP, May 1992.

What We've Learned About Assessing Hands-On Science

Assessing scientific inquiry is more complex than political rhetoric pushing performance tasks indicates, this team of scientists found. And, unless carefully crafted and blended into science instruction, assessments alone are unlikely to boost achievement.

RICHARD J. SHAVELSON AND GAIL P. BAXTER

Over the past three years, our team of researchers, scientists, and science teachers at the University of California, Santa Barbara, the California Institute of Technology, and the Pasadena Unified School District has sought to create assessments that support good science teaching at the elementary level (Baxter et al. in press; Shavelson et al. 1991). In particular, our goal has been to develop activities that permit students to pursue an experimental inquiry focusing on process skills (such as observing and inferring) and construction of new knowledge (such as understanding the effects of insulators on electric current). In our definition, assessments consistent with good teaching invite students to perform concrete, meaningful tasks such as a laboratory experiment to determine, for example, how certain kinds of insects respond to changes in environment. Scoring of performance focuses on the reasonableness of the procedure used to carry out the investigation, not just on the "right answer."

In our work, we assumed that the ideal assessment would be direct observation of a student pursuing a scientific inquiry with laboratory equipment and materials. This observation would be made by scientists and science teachers trained to score performance in real time. That is, the ideal was predicated on the assumption of the symmetry of teaching and testing: an ideal assessment would be a good teaching activity and, indeed, might even serve as a teaching activity when not used for assessment.

However, observations of individual student performance are costly, time consuming, and difficult to obtain. With the ideal performance assessment as a benchmark, we developed and evaluated alternatives (or surrogates) to the benchmark. They were, in order of decreasing verisimilitude, (1) lab notebooks in which students recorded their procedures and conclusions; (2) computer simulations of the hands-on investigations; (3) short-answer paper-and-pencil problems in planning, analyzing, and/or interpreting experiments; and (4) multiple-choice items developed from observations of students conducting hands-on investigations. Finally, we compared the benchmark and surrogate assessments to a traditional multiple-choice science achievement test, the Comprehensive Test of Basic Skills (CTBS).

We developed and collected data using three hands-on investigations:

- Paper Towels — Given laboratory equipment, conduct an experiment to determine which of three different paper towels soaks up the most water.
- Electric Mysteries — Given six "mystery boxes," determine their contents by connecting electric circuits to them.
- Bugs — Determine sow bugs' preferences for various environments (for example, dark or light, dry or wet).

The performance of more than 300 students, some experienced in hands-on science and some who had received minimal science instruction from a textbook on health, was observed and scored in real time by science educators. In addition, all students completed corresponding notebooks, computer simulations, paper-and-pencil measures, and the CTBS.

Four questions guided our research: (1) Could reliable measures of hands-on performance and of surrogate assessments be developed? We wanted these measures to permit a wide variety of student responses found when doing science. We also wanted to develop a method to score performance that captured the diversity of procedures and put them on a common scale. (2) Could the performance of students with different instructional experiences (hands-on vs. textbook) be distinguished? We expected students experienced with hands-on science to perform better on the

benchmark and closest alternatives than students in a textbook curriculum program. (3) Do the performance assessments provide information about student achievement not available from traditional multiple-choice science tests? If not, perhaps nothing new had been developed. (4) Do the surrogate assessments capture the information gained from the benchmark? If so, then dollars and time can be saved in administration and scoring.

Hands-On Investigations

Students conducted the three investigations in approximately 1 1/2 hours while being observed by a scientist or science educator trained to score student performance.

Paper Towels. Students used a laboratory setup to determine which of three paper towels held the most and least water. Students were told that they could use all or some of the equipment, whatever they needed. A scheme was developed to score the diversity of procedures used to carry out the investigation on a common scale (see fig. 1). An outstanding investigation completely saturated each towel, determined the amount of water each held by a method that was consistent with the way the towel was wetted, and the entire procedure was done carefully. For example, a student might saturate the towel in the pitcher of water and weigh it in the scale, carefully removing the excess water in the scale after weighing each towel. Carelessness, inconsistencies in the method of wetting the towel and measuring the results, incomplete saturation, and irrelevant methods resulted in less than outstanding scores. The scoring scheme identified the procedure used and could thereby characterize performance in terms of both processes and outcomes. Moreover, several different performances

High-quality assessments throughout a course are vital if teachers are to accurately appraise performance.

FIGURE 1

PAPER TOWELS INVESTIGATION—HANDS-ON SCORE FORM

Student_____ Observer_____ Score _____ Script_____

1. **Method** A. Container: Pour water in/put towel in; Put towel in/pour water in; 1 pitcher or 3 beakers/glasses. B. Drops. C. Tray (surface): towel on tray/pour water on; pour water on tray/wipe up

2. **Saturation** A. Yes B. No C. Controlled

3. **Determine Result**
   - A. Weigh towel
   - B. Squeeze towel/measure water (weight or volume)
   - C. Measure water in/out
   - D. Time to soak up water
   - E. No measurement
   - F. Count # drops until saturated
   - G. See how far drops spread out
   - H. Other _____

4. **Care in Measuring** Yes No

5. **Correct Result** Yes No

Grade	Method	Saturate	Determine Result	Care in Measuring	Correct Answer
A	Yes	Yes	Yes	Yes	Yes
B	Yes	Yes	Yes	No	Yes/No
C	Yes	Yes/Controlled	Error		Yes/No
D	Yes	No	Missing		Yes/No
F	----------	----------	No Attempt	----------	----------

BEST COPY AVAILABLE

could result in the same letter grade.

Bugs. Students used laboratory apparatus to determine the preferences of sow bugs for environments that were light or dark, damp or dry, or some combination of the four. The scoring scheme resembled the one used in the towels investigation.

Electric Mysteries. This investigation was a bit different from the others. Students were given batteries, bulbs, and wires and asked to determine the circuit components hidden in a set of mystery boxes (see fig. 2). Performance was scored on the basis of (1) their determination of the contents of each box and (2) the sequence of tests they conducted to determine the contents.

Notebook Surrogates

For each investigation students were asked to keep notebooks describing their investigation so that a friend could repeat it exactly. Notebooks have several advantages for large-scale assessment. Through the use of notebooks instead of expert observers, many students can be tested with hands-on investigations. Notebooks provide an opportunity for students to express themselves in writing, an important skill in science and a way of integrating curricular areas. And the notebooks can be scored quickly — in about one to two minutes per student. Notebooks, then, preserve much of the hands-on investigation while reducing time and cost of expert observers. They also capture the rather inventive nature of the investigations and ways of reporting on them.

Computer Simulation Surrogates

We developed our own computer simulation for the Electric Mysteries and Bugs investigations to replicate, as nearly as possible, the hands-on investigations. (The Paper Towels investigation could not be simulated adequately.) For the electric circuits investigation, students used a Macintosh computer with a mouse to connect circuits to the mystery boxes to determine their contents. The intensity of the luminosity of the bulb in a real external circuit was accurately simulated. Students could connect a multitude of circuits if they so desired. Alternatively, they could leave a completed circuit on the screen for comparison. A teacher-directed tutorial prior to the test provided students with instructions on how to record their answers, erase wires, save their work, or look at a previous page of their work on the screen. The computer recorded every move the student made. The bug simulation was constructed similarly, though it was not possible to record every move the student (and the bugs) made.

Computer simulations have a number of desirable properties for assessment. They are less costly and time consuming to administer than hands-on assessments, though development costs are considerable. Students can be tested in groups by a parent or volunteer who has been briefed on how the simulations work. Student performance can be scored quickly and easily using the scoring system developed for the hands-on investigations. In addition, a computer simulation maintains a full record of performance, so that teachers and/or students can review problem-solving processes. Finally, students experiment with the technology, discovering solutions to problems that they might not have found with other types of assessments. In other words, the computer assessment provides a good instructional tool.

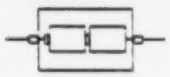
Pencil-and-Paper Surrogates

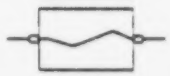
Short-answer and multiple-choice questions were chosen to parallel, in content, the three hands-on investiga-

FIGURE 2

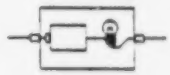
HANDS-ON ELECTRIC MYSTERIES INVESTIGATION

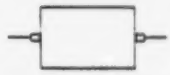
Find out what is in the six mystery boxes A, B, C, D, E, and F. They have five different things inside, shown below. Two of the boxes will have the same thing. All of the others will have something different inside.

Two batteries: 

A wire: 

A bulb:

A battery and a bulb: 

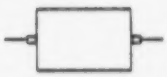
Nothing at all: 

You can use your bulbs, batteries, and wires any way you like. Connect them in a circuit to help you figure out what is inside.

When you find out what is in a box, fill in the spaces on the following pages.

Box A: Has _____ inside.

Draw a picture of the circuit that told you what was inside **Box A**.



How could you tell from your circuit what was inside **Box A**?

_____

_____

_____

Do the same for **Boxes B, C, D, E, and F**.

_____

_____

tions. For the Electric Mysteries short-answer questions, students received a pictorial representation of a problem similar to one encountered during the hands-on investigation. For example, students might be asked how they would determine the contents of a particular mystery box without looking inside. For the Paper Towels and Bugs questions, students were given descriptions of portions of the investigations and questioned about the control of variables, the setup of the experiment, or the best method to use for measuring results.

Multiple-choice questions began much like the short-answer questions. Rather than formulate a response, students chose among four alternatives, all of which were based on observed performance. For example, an Electric Mysteries question presented alternative circuits connected to a box and asked students to indicate which circuit would tell them what was inside the box.

The paper-and-pencil surrogate assessments differ fundamentally from the other surrogates; they do not respond to the actions taken by the students. Even if a paper-and-pencil test provided immediate written feedback on a decision made by a student (Shulman and Elstein 1975), we doubt it would have the same impact as the feedback from the real-life (hands-on) or lifelike (computer) assessments. We may not be able to develop paper-and-pencil surrogates that overcome this limitation.

Findings

We found the process of creating performance assessments symmetric with good teaching activities to be time consuming, requiring considerable scientific and technological know-how. Development of quality performance assessments requires multiple iterations through a sequence of development, tryouts with students (getting their thoughts and comments), and revision. Short-circuiting this process leads to ill-conceived and poorly constructed assessments. Such assessments are as likely to lead to poor teaching — if teachers teach to the test, and they do (Smith 1991) — as are ill-conceived and poorly executed classroom activities.

If teachers teach to poorly constructed assessments, these assessments are likely to distort good hands-on science teaching.

Once the performance assessments were constructed, our research posed four questions about them: (1) Can they provide reliable measurements? (2) Are they sensitive to students' instructional experiences? (3) Do they provide achievement information that differs from traditional measures? (4) Are they interchangeable? Our findings (some good news, some bad news) balance the political rhetoric pushing implementation of performance assessments with a cold reality.

Reliability. Raters can reliably evaluate students' hands-on performance on complex tasks in real time. Reliabilities are high enough (above 0.80) that a single rater can provide a reliable score. But task-sampling variability is considerable. Some students perform well on one investigation while others perform well on a different investigation: general "expertise" is more in the mind of the beholder than in performance. To get an accurate picture of individual student science achievement, the student must perform a substantial number of investigations — perhaps between 10 and 20.

Instructional history. Performance assessments can distinguish students with different instructional histories. Assessments that are closely linked to a specific domain of knowledge (for example, electric circuits/electricity) are more sensitive to performance differences than more general process assessments (for example, the Paper Towel investigation). But to be sensitive to instructional history, performance assessments must be carefully crafted to measure more than science process. They need to measure the application of both concepts and procedures.

Relation to multiple-choice tests. The good news is that performance assessments do *not* duplicate information about student achievement provided by traditional tests (average correlation is about 0.45). They tap somewhat different aspects of achievement. But we are not sure what aspects of achievement multiple-choice tests or performance assessments do and do not tap. Indeed, a combination of indicators (multiple-choice, performance assessments, and others) may be needed to provide a complete picture of achievement.

Interchangeability of surrogates. Certain surrogate assessments appear to be interchangeable with their corresponding benchmark. This is especially true of notebooks for student-level assessment. Computer simulations are interchangeable with their corresponding benchmarks if the intent is to estimate *classroom-level* mean performance. But for individual students, measures of science achievement are highly sensitive not only to the investigation used (for example, Bugs vs. Electric Mysteries), but also to the method used to measure performance (for example, observation vs. simulation). Some students' scores depend on the particular investigation

(Electric Mysteries, Paper Towels, Bugs) *and* on the particular method used to assess performance (observation, notebook, computer simulation, paper-and-pencil). Indeed, each combination of investigation and method provides different insight into what students know and can do.

Conclusions. A fundamental assumption made by policymakers (for example, Bush 1991) and other education reformers (for example, Wiggins 1989) is that, by changing the nature of the achievement test, teachers who teach to the test will have to change and improve their teaching. Our experience with performance assessments suggests that this assumption is, at best, half true. Teachers will indeed change the way they teach if held accountable by performance assessments. But, without high quality assessments and staff development in quality instruction, they very well may not improve their teaching. Moreover, one-shot, end-of-year tests cannot provide adequate information on individual-level student achievement. Continuous assessment throughout the course of instruction is needed to accurately reflect student science achievement.

Assessment Development

Performance assessments are very delicate instruments. They need to be carefully crafted, each requiring a specially developed or adapted scoring method. Shortcuts taken in developing these assessments will likely produce poor measuring devices. If these instruments are used to judge the quality of education in classrooms — and they will be used for that purpose — then teachers will teach to the test. If teachers teach to poorly constructed assessments, these assessments are likely to distort good hands-on science teaching.

This conclusion was brought home to us in our observations of someone we consider to be an outstanding hands-on science teacher. In her class, a unit on electricity is taught in a series of lessons. Small groups of students conduct hands-on investigations, not unlike the ones we have developed. Students keep notebooks and draw conclusions based on the outcomes of their experiments and their discussions with other groups of "scientists" in the class.

Excited about hands-on science teaching, our teacher volunteered her class for pilot testing California's new hands-on science assessments. These assessments were constructed under severe time and cost constraints, and consequently involved minimal trials with students. To meet testing time and space constraints, and in recognition of differing curriculums in the state's elementary schools, the assessments were accompanied by detailed directions to students. Students read the instructions, followed explicit procedures, and reported what they found in the spaces provided. The assessments were more like recipes (first do this step, then do this step) than like scientific investigations.

Based on her experience with these assessments, our exemplary teacher began to modify her teaching to correspond to the state's pilot assessments. Rather than open-ended, group problem solving, emphasis was placed on reading instructions and carefully following format, carrying out a set of required procedures, and recording findings in a prespecified format in notebooks. (For example, students were admonished to "be sure to write complete sentences or the state will grade you down.") The essence of *doing science* was becoming one of following procedures. The story ends happily. After we pointed out what was happening, the teacher went back to her "old ways."

Staff Development

Unless provided the scientific and pedagogical knowledge required for hands-on science teaching, teachers may very well flounder in their attempts to match their teaching to the testing. Once again, an anecdote may bring home this conclusion. One of the teachers whose class was participating in our research knew of the Paper Towels investigation. Teaching to the test, she instructed her students to saturate the towels completely, using the timer to ensure that each towel was saturated for at least 10 minutes — a total of 30 minutes for saturation! In reality, the towels could be saturated in a matter of seconds. This led her students, when tested, to perform in a clearly stylized manner. She had informed the students, perhaps unintentionally, that science is a set of precise steps that must be carried out invariably, regardless of whether they make sense. Her approach was not particularly conducive to scientific exploration. Although the teacher could teach the students how to "do" the experiment, what was missing was an understanding of the essence of doing science.

Curriculum-Embedded Assessments

To obtain sufficiently large samples of student performance, assessments may need to be taken throughout the academic year. For example, students might receive the Electric Mysteries assessment embedded in the activities composing a unit on electricity. Likewise, assessments would be embedded in three or four other units as well.

The embedded assessments have several desirable characteristics. They provide almost immediate feedback to teachers on their students' perfor-

mance, and on how this performance compares with that of students in comparison schools. Moreover, the assessments reinforce hands-on teaching and learning. Nowhere is the symmetry between teaching and assessment more apparent than with embedded assessments.

Embedded assessments do not preclude an end-of-year examination. The latter provides both additional information on achievement and an external audit ensuring data credibility to the various audiences interested in educational accountability.

Will Achievement Improve?

Results of our research suggest that the political rhetoric calling for immediate reform of national, state, and local testing systems far exceeds current technological capability and ignores educational and social consequences. No doubt assessment systems will be changed in the very near future. The politicians will have their day. We suspect that the initial impact will be to change classroom activities and the nature of assessment, possibly embedding assessments in classroom activities. However, without quality instrumentation and extensive staff development, the bottom line — achievement — will probably not change.

The nation may be capable of producing the kind of assessment systems currently envisioned by the rhetoric if the politicians stick to their guns and do not retreat to the usual multiple-choice testing. Politicians need to provide resources for preparing beginning and current teachers for teaching and testing reforms, and for fine-tuning the assessments through research, social debate, and revision. With the symmetry between assessment and teaching firmly established, the bottom line — achievement — may very well improve. □

Authors' note: This research was supported by a grant from the National Science Foundation (No. SPA-8751511). The ideas presented reflect those of the authors and not necessarily the NSF.

References

Baxter, G. P., R. J. Shavelson, S. R. Goldman, and J. Pine. (In press). "Procedure-Based Scoring for Hands-On Science Assessments." *Journal of Educational Measurement.*

Bush, G. W. (1991). *America 2000: An Education Strategy.* Washington, D.C.: U.S. Department of Education.

Shavelson, R. J., G. P. Baxter, J. Pine, and J. Yure. (1991). "Alternative Technologies for Assessing Science Understanding." Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shulman, L. S., and A. S. Elstein. (1975). "Studies of Problem Solving, Judgment, and Decision Making: Implications for Educational Research." In *Review of Research in Education* 3: 3-42, edited by F. N. Kerlinger.

Smith, M. L. (1991). "Put to the Test: The Effects of External Testing on Teachers." *Educational Researcher* 20, 5: 8-11.

Wiggins, G. (1989). "A True Test: Toward More Authentic and Equitable Assessment." *Phi Delta Kappan* 70, 9: 703-713.

Richard J. Shavelson is Dean of the Graduate School of Education and Professor of Research Methods. **Gail P. Baxter** is a researcher in the Graduate School of Education and co-principal investigator on two National Science Foundation grants. They can be contacted at the University of California, Santa Barbara, CA 93106.

SECTION III: Math Assessment K-6

Reprinted with permission from EDUCATIONAL LEADERSHIP, February 1993.

NCTM's *Standards*: A Rallying Flag For Mathematics Teachers

Thomas A. Romberg

The development of the *Standards* is the story of a professional community recognizing a need, reaching consensus through review and analysis, and conceiving a clear vision and a strategy for mathematics reform.

Curriculum and Evaluation Standards for School Mathematics has been hailed as the exemplar of what is needed in all curricular areas if we are to reform American education during the coming decade (National Council of Teachers of Mathematics 1989). As chair of the commission that produced this 258-page document, my intent here is to provide an account of how the mathematical sciences education community developed the *Standards*.

Origin of the *Standards*

The strategy we used came about as a result of two conferences held in 1983 in response to the perceived crisis in education, as voiced in *A Nation at Risk* and *Educating Americans for the Twenty-First Century*.[1] The first conference was sponsored by the Conference Board of the Mathematical Sciences and funded by the National Science Foundation; the other was jointly sponsored by the National Council of Teachers of Mathematics (NCTM) and the University of Wisconsin and funded by the U.S. Department of Education. The dozen or so recommendations contained in the reports of both meetings were strikingly similar.[2] Each suggested the development of a "new content framework" for the K-14 mathematics curriculum.

In 1986 the National Council of Teachers of Mathematics organized the Commission on Standards for School Mathematics to prepare a draft document. Composed of classroom teachers, supervisors, mathematics educators, and mathematicians, the Commission formed four working groups, one each to address curriculum for grades K–4, 5-8, 9-12, and a group to address evaluation. In developing the draft, the groups consulted reports and background papers related to the calls for reform; a variety of research reports;[3] recent state curriculum documents (California, Oregon, Wisconsin); curriculum documents from other countries (Australia, The Netherlands, Japan, The United Kingdom); and two summary papers outlining the perspectives and the tasks to which we were committed.

During the summer of 1987, the groups prepared and distributed 10,000 copies of our draft for review by NCTM and the Mathematical Sciences Education Board. We received more than 2,000 suggestions from parents, business leaders, teachers, and mathematicians. We also asked other professional mathematics organizations to critically review and endorse the document. In 1988, the writing groups met again to revise the document, which, after additional review and editing, was published in 1989. Also in 1988, NCTM created several more working groups to develop the Professional Teaching Standards, which were published in 1991. Three facts are important in this sequence of events:

1. The *Standards* were created as a consequence of scholarly review and analysis.

2. Review and consensus involving many groups were paramount in the process.

3. Using its own resources, a professional organization produced the document. No authority to develop standards was demanded by a governmental agency or corporate group. The professional community decided it was needed and took the initiative.

Nearly half of American students never study any mathematics beyond arithmetic. Here, a varsity athlete practices calculus.

The Reform Vision

The vision of school mathematics expressed in the *Standards* is captured in this statement: "All students need to learn more, and often different, mathematics. . . . Instruction in mathematics must be significantly revised" (NCTM 1989, p. 1). This statement has five implications:

1. Teaching mathematics to "all students" emphasizes that anyone who is to be a productive citizen in the 21st century must be mathematically literate—including not only "talented white males" but all underrepresented groups.

2. "More mathematics" implies that all students need to learn more than how to manipulate arithmetic routines. At present, nearly half of American students never study any mathematics beyond arithmetic.

3. "Often different mathematics" indicates that all students need to learn concepts from algebra, geometry, trigonometry, statistics, probability, discrete mathematics, and even calculus.

4. "To learn" means more than to memorize and repeat. Learning involves investigating, formulating, representing, reasoning, and using appropriate strategies to solve problems, and then reflecting on how mathematics is being used.

5. "Revised" instruction implies that teachers and students need to envision mathematics classrooms as discourse communities where conjectures are made, arguments presented, strategies discussed, and so forth.

The content that all students should have an opportunity to learn is described in detail in *Curriculum and Evaluation Standards for School Mathematics*. To capture the importance of both technical and reflective knowledge, we chose the term *mathematical power*, which envisions a citizenry and society *empowered* by mathematics. For the individual, mathematical power means having the experience and understanding to participate constructively in society. For a culture to be mathematically powerful, its citizens must have the understanding and experience to undertake the routine tasks of everyday life, to operate as a democratic society, and to progress as a civilization. This means the acquisition, in society, of a critical mass of understanding, experience, and attitude, in addition to a substantial range of special expertise.

Thus, students need to learn to value mathematics, to reason and communicate mathematically, and to become confident in their power to use mathematics coherently to make sense of problematic situations in the world around them. Hence, the *Standards* advocate the application of four basic standards to all of the content standards proposed: mathematics as problem solving, mathematics as reasoning, mathematics as communication, and mathematical connections (making linkages within mathematics and between mathematics and the real world).

The Reform Strategy

The Commission's intent to develop a document titled *Standards* was deliberate. According to the *Oxford American Dictionary*, a *standard* is "a thing or quality or specification by which something can be tested or measured," "the required level of quality," or "a distinctive flag" (1980, p. 666). For us a standard specified criteria to judge the quality of the mathematics

BEST COPY AVAILABLE

curriculum and methods of evaluating it. In order to judge quality, we decided that there must be agreed-upon expectations. Thus, standards become statements about what is valued.

Historically, groups have adopted standards to (1) assure quality, (2) indicate goals, and (3) promote change. The NCTM Commission on Standards was created to achieve each of these three objectives. In the United States, any group or individual can produce and sell a mathematics textbook, test, or curriculum guide. Although most such efforts are well intentioned, the recent proliferation of manipulatives, software, and modules have made minimal curricular standards a necessity. The Commission focused on the development of standards as statements of expectations or as "criteria for excellence." However, part of our broader strategy was to provide mathematics teachers "a distinctive flag" to justify their demands for change.

The need for such a document was made explicit at a 1983 conference of publishers, held at the University of Wisconsin. Since we live in a supply-and-demand economy, the publishers argued, if the mathematics community wanted different texts and tests, a demand would have to be created. To respond to this challenge, the mathematical sciences education community has followed a seven-step iterative strategy to reform school mathematics. The steps and the relationships between them are as follows:

1. Before any plan can proceed, a *need* for change must be established. First, evidence must be presented that the current system is not effective. Second, planners must decide whether the problem is the result of a lapse in quality control, a design flaw, or a combination of both (Cuban 1988). Third, if a new design is needed, planners must be aware of the traditions that have to be challenged. The need for a new blueprint for mathematics education has been well documented.[4]

2. *Vision* is a key factor if a new design is needed. Design change involves not only eliminating or replacing inadequate materials or practices, but creating a new system. Realizing a vision necessarily includes consideration of values, goals, and standards. The *Standards* were designed to fulfill this vision.

In economics, a product is judged of good quality if it satisfies customer needs while making the company a profit.

3. *Planning* includes involving everybody in a system or school in arriving at consensus about the details of long- and short-range plans (with timetables) for change. It is at this step that demand is created.

4. The next step involves identifying specific *elements* of the system to be targeted for change (curriculum materials, instructional methods, examinations, teachers, technology) and prioritizing them.

5. Any system depends on *suppliers*. In a supply-and-demand society, schools must demand that suppliers (textbook publishers, testing companies, staff developers, university teacher education programs) contribute the ingredients necessary for the desired changes in elements.

6. Then it is time to *operationalize* the new materials, procedures, and programs. Feedback from this trial phase is matched with the vision and the plan for judgment, and revisions may be made.

7. Finally, a *product* (a curriculum, an instructional procedure, assessment materials) is developed. In economics, a product is judged of good quality if it satisfies customer needs while making the company a profit. In education, quality should be judged in terms of what students are able to do and whether this meets society's needs. Again, via feedback, we now return to the vision and our goals and objectives to update or revise the vision, plans, specific elements, and so forth.

Our intent is that school staffs at all levels use this seven-step strategy to develop a reform curriculum for K-12 mathematics. In particular, each mathematics teacher must understand the need for change, own the vision, participate in planning, become a spokesperson for the demands for new products and processes, try out new materials, and judge student progress toward the reform vision. Teachers need to understand the critical role they play in curriculum reform efforts.

Status of the *Standards*

Implementation of the *Standards* means a number of things need to occur: using the documents to plan change, making demands of suppliers, enhancing the professional status of mathematics teachers and educators, and empowering teachers to voice their views. Planning for change in education takes place on at least three organizational levels.

Reform initiatives at the *national* level are being influenced by the *Standards*. They were a focal point in the Bush administration's strategy for school reform and were adopted by the National Education Goals Panel. At the *state* level, reform plans should go beyond goal statements and general expectations to curricular frameworks for schools. Several states (California and Texas, for example) have recently produced frameworks based on the *Standards*, and several others are in the process of developing them. In addition, many states (19, at last count) have initiated new assessment practices, in part to align testing with the goals of the reform effort.

At the *local district* level, plans for change should go beyond specifying materials and resources to designating how teachers and students should interact with them. It is encouraging that many districts have formed

curriculum committees; however, it is hard to estimate how many of these districts have developed implementation plans.

The evidence that the demands for change are being heard is clear. The National Science Foundation is funding a variety of curriculum and teacher enhancement projects. The Department of Education is encouraging the use of Eisenhower funds to assist teachers in developing new materials. California, Texas, and other states are urging publishers to develop new texts to meet their frameworks. In fact, at current professional meetings, book publishers are claiming either that their current materials meet the criteria set forth in the *Standards* or that they are developing new materials to meet them. Test publishers, too, are busy developing new instruments to be aligned with the *Standards*.

The professional status of mathematics teachers and educators and their empowerment has also been enhanced. The creation of the Mathematical Sciences Education Board (another of the recommendations) has provided the mathematics community with a national agency to represent its constituents. Membership in national and local mathematics teachers' organizations is at an all-time high, and attendance at meetings has set records every year. Finally, an increasing number of mathematics teachers are being asked to testify at hearings and serve on national and state committees dealing with mathematics education.

The impact of the *Standards* document since its publication three years ago has gone far beyond our expectations. However, the hard part of systemic change is yet to happen. Only when the plans and new products are being used in classrooms by teachers who own the philosophy and vision expressed in the *Standards* will real change begin to occur.

Challenges That Remain

To carry out the vision of a reformed school mathematics curriculum, policymakers and school staffs face a number of issues and problems.

In education, quality should be judged in terms of what students are able to do and whether this meets society's needs.

The meaning of "standards"? To many persons, the term *standards* implies measuring student performance. Although not a serious issue, the use of the word to indicate a vision that mathematics teachers can rally around is unusual and requires explanation.

Nominal change. To appease demands for change, producers often change labels but not substance. For example, since 1990 the authors of the National Assessment for Educational Progress have claimed substantial changes in items in the test battery in light of the reform expectations. In fact, both the 1990 and 1992 batteries show little alignment with the NCTM *Standards* (Romberg et al. 1992).

What is mathematics? "Most of the population perceive mathematics as a fixed body of knowledge long set into final form. Its subject matter is the manipulation of numbers and the proving of geometrical deductions. It is a cold and austere discipline which provides no scope for judgment or creativity" (Barbeau, 1989, p. 2). The aims of teaching mathematics need to go beyond this narrow view to empower learners to create their own mathematical knowledge; to reshape mathematics, at least in school, to give all groups more access to its concepts and to the power its knowledge brings; and to bring the social contexts of the uses and practices of mathematics into the classroom. When mathematics is seen in this way, it can then be studied in living contexts that are meaningful to learners, including their languages, cultures, and everyday lives, as well as their school-based experiences.

The "saber-tooth tiger" content. Topics should be included because of their inherent worth, not because they have "always been part of the curriculum." Peddiwell's (1939) satirical tale of continuing to teach students techniques to scare saber-tooth tigers with fire long after the tigers have become extinct is analogous to the continued expectation that students master interpolation of logarithms, square roots, long division, and myriad other routine procedures long after computers have automated them. Further, room needs to be made in the curriculum for new, or newly important, topics such as statistics or discrete mathematics.

Teacher independence and isolation. One tradition of schooling is that teaching happens behind closed doors (Metz 1978). Such independence allows teachers to take risks and be creative. Taken to an extreme, independence often leads to isolation. Porter (1988) refers to one consequence of independence as a curriculum "out of balance." Elementary school mathematics is routinely taught in such a way that students are repeatedly exposed to the same content from one grade to the next, with skills typically receiving 10 times the emphasis given to either conceptual understanding or application. The reform vision sees as the norm a balanced curriculum arrived at via teacher collaboration, joint planning of lessons, and shared judgments about student performance.

The "hidden" curriculum. Probably the biggest challenge of reform for parents, administrators, and even teachers is its threat to school routines that are based upon existing architecture, organization, and management. Reform ideas are challenging such notions as:

- "Math is what we do to quiet kids down after recess."
- "Drill on procedures teaches students how to follow rules."

Only when the plans and new products are being used in classrooms by teachers who own the philosophy and vision expressed in the *Standards* will real change begin to occur.

- "Success on a math test is essential for tracking."
- "Math should not be fun. It teaches students about drudgery and failure."
- "Curricular changes must improve standardized test scores."

Isn't this a repeat of the "new math"? This is a natural question, given the failure of post-Sputnik attempts to develop a new mathematics curriculum. The answer is: NO! The "new math" was an elitist attempt by university mathematicians to better prepare college-bound mathematics students for a changed collegiate curriculum (Romberg 1990). The current movement focuses on mathematics for all students and is being organized by the teachers of mathematics at all levels.

How is this effort related to other reforms? Although the objectives of the mathematical sciences education community were developed independently of many other programs, if they embrace the same assumptions about subject matter content and pedagogy as other change plans, then they should be compatible. Systemic change based on the same ideas is needed.

A Vision and a Strategy

There is no question that the current school mathematics curriculum is out-of-date. The majority of American students do not have an opportunity to become empowered mathematically even though our culture demands mathematical literacy of its citizens. Creating new curriculums will take time, hard work, commitment, patience, and persistence.

In *Curriculum and Evaluation Standards for School Mathematics*, the mathematical sciences education community has proposed both a vision of a mathematics curriculum designed to meet the need and a strategy that districts and schools can follow to construct a curriculum consistent with this vision. For teachers, the *Standards* provide a flag to rally around as they fulfill their essential role in mathematics reform. ■

---

[1]National Commission on Excellence in Education. (1983). *A Nation At Risk: The Imperative for Educational Reform*. (Washington, D.C.: U.S. Government Printing Office); and National Science Board Commission on Precollege Education in Mathematics, Science, and Technology. (1983). *Educating Americans for the Twenty-First Century: A Plan of Action for Improving the Mathematics, Science, and Technology Education for All American Elementary and Secondary Students So That Their Achievement Is the Best in the World by 1995*. (Washington, D.C.: U.S. Government Printing Office).

[2]The two reports are: *New Goals for Mathematical Sciences Education*, report of a 1983 conference sponsored by the Conference Board of the Mathematical Sciences, Airlie House, Warrenton, Va., (Washington, D.C.: CBMS); and T.A. Romberg. (1984). *School Mathematics: Options for the 1990s*. Chairman's Report of a Conference. (Washington, D.C.: U.S. Government Printing Office).

[3]Summarized in a three-volume report: T.A. Romberg and D.M. Stewart. (1987). *Monitoring of School Mathematics: Background Papers* (Madison: Wisconsin Center for Education Research).

[4]In particular, see *Everybody Counts*. (1989). (Washington, D.C.: Mathematical Sciences Education Board).

References

Barbeau, E. J. (1989). *Mathematics for the Public*. Paper presented at the meeting of the International Commission on Mathematical Instruction, Leeds University, Leeds, England.

Cuban, L. (1988). "A Fundamental Puzzle of School Reform." *Phi Delta Kappan* 69, 5: 341-344.

Metz, M. H. (1978). *Classrooms and Corridors*. Berkeley: University of California Press.

National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, Va.: NCTM.

Oxford American Dictionary. (1980). New York: Oxford University Press.

Peddiwell, J. A. (1939). *The Saber-Tooth Curriculum*. New York: McGraw-Hill.

Porter, A. (1988). *A Curriculum Out of Balance: The Case of Elementary School Mathematics* (Research Series No. 191). East Lansing, Mich.: Michigan State University, Institute for Research on Teaching.

Romberg, T. A. (1990). "New Math" Was a Failure—Or Was It? *UME Trends* 2, 6: 1-3.

Romberg, T. A., M. Smith, S. Smith, and L. Wilson. (1992). *The Feasibility of Using International Data to Set Achievement Levels for the National Assessment of Educational Progress (NAEP)*. Madison, Wis.: National Center for Research in Mathematical Sciences Education.

Thomas A. Romberg is Sears Roebuck Foundation-Bascom Professor in Education and Director, National Center for Research in Mathematical Sciences Education, Wisconsin Center for Education Research, at the University of Wisconsin-Madison, 1025 W. Johnson St., Madison, WI 53706.

Measuring What's Worth Learning

Reprinted with p
Measuring Up: P
Mathematics Asses
right 1993 by the
Academy of Science
of the National Ac
Washington, DC.

The spotlight of educational reform continues to sweep across the stage of mathematics. First curriculum, then teaching, and now assessment have come under intense professional and public scrutiny. Amid deteriorating public confidence in the quality of American education, the mathematical community is addressing multiple challenges to articulate and implement effective standards in the key arena of testing, assessment, and accountability.

In the center of the assessment stage are three elements contesting for leadership. Conventional testing offers comfortable short-response tests on traditional content that are taken by millions of students every year. Reformers, including authors of the two K-12 *Standards* documents from the National Council of Teachers of Mathematics (NCTM), call for fundamental change — different in content, in format, and particularly in spirit. To this well-rehearsed contest of traditionalist vs. reformist has now been added a third movement arriving from outside the educational community: the political call for assessment of progress towards our nation's new standards in mathematics education.

In the decade since publication of *A Nation at Risk*, the United States has moved a long way toward a new consensus for education. Talk of national standards, once taboo, is now commonplace; so too is talk of alternative school structures

ssion from
types for
ent. Copy-
ational
. Courtesy
demy Press,

and innovative licensure for teachers. It is now time to develop a new national understanding of standards-based performance as the true measure of educational progress.

Throughout this decade, mathematics has led the way in educational reform. The 1989 MSEB publication *Everybody Counts* was followed in just two months by publication of the NCTM *Curriculum and Evaluation Standards for School Mathematics,* with its theme of developing mathematical power in all students. Undergirding these reports are three fundamental principles of testing, assessment, and accountability:

- Tests should measure what's worth learning, not just what's easy to measure.
- Progress depends on constant correction based on feedback from assessment.
- Schools are accountable, both to taxpayers and to students.

Even as the renewed public scrutiny compels educators to demonstrate that children are learning, the NCTM's *Standards* require new ways of measuring what is being learned. Because the linkage between tests and teaching is so close, it is vitally important for the United States that assessment be based on instruments that are properly aligned with the goals of the *Standards.*

The Challenge

At the National Summit on Mathematics Assessment in April 1991, Governor Roy Romer, in his capacity as Co-chair of the National Education Goals Panel, challenged the mathematical community to show the nation what mathematics educators mean by mathematical power and what new and more demanding standards will mean for our young people. One month later, the MSEB authorized creation of *prototypes* of tasks that could be used to assess fourth-graders' mathematical skills and knowledge, thereby providing examples of what children educated according to the new standards should be able to do. They wanted to be sure that the voice of mathematics was heard early and clearly in the assessment reform movement. The MSEB determined that it should be prepared to show, by

Why we are doing this

example, the type of assessment exercises that would be appropriate to measure our nation's progress toward the goals of mathematics education.

To create the prototypes, the MSEB subsequently convened a small writing group of mathematics educators, teachers, and mathematicians. Taking up Governor Romer's challenge, the writing group created a sampler of tasks to encompass many of the goals for mathematics instruction that are expressed in the NCTM *Standards*. These tasks, which illustrate what a standards-based education really means, have been pilot tested on a limited basis in four states. Many have been revised, often more than once, but all can benefit from continued improvement and adaptations.

Readers who skip ahead will see that these prototypes are not only innovative and challenging but also just plain fun. Teachers, children, and even parents should find these tasks both engaging and surprising. We invite readers to try them, either before or after reading the surrounding analysis.

The Criteria

Not surprisingly, the MSEB writing group debated extensively the criteria for prototypical assessment tasks. They faced the pioneer's challenge — to use incomplete information as a basis for decisions whose consequences are difficult to foresee. From these discussions emerged several criteria that helped shape the nature and selection of prototypes in this volume:

What we are trying to do

- *Mathematical content:* The tasks should reflect the "spirit" of the reform movement, but not necessarily be limited by particular curricular content, present or planned. Many of the tasks should incorporate a variety of mathematics, particularly in areas such as statistics, geometry, and probability that are least often emphasized in traditional K-4 programs.
- *Mathematical connections:* Everyone involved in the mathematics reform movement, from classroom teach-

ers to national policy makers, agrees on the importance of connecting mathematics — to science, to social science, to art, to everyday life, and to other parts of mathematics. Accordingly, the prototypes should develop links with science, with the visual arts, and with the language arts.

- *Thoughtful approaches:* Insofar as possible, the tasks should promote "higher-order" thinking. Just as the verbs explore, justify, represent, solve, construct, discuss, use, investigate, describe, develop, and predict are used in the *Standards* to convey "active physical and mental involvement of children" in learning mathematics, they are appropriate to seek in assessment activities as well. Further, given a choice between cognitive complexity and simplicity, the focus of these tasks should be on the former.

- *Mathematical communication:* The tasks should emphasize the importance of communicating results — not simply isolated answers, but mathematical representations and chains of reasoning. Children should have opportunities to work in groups to explain their thinking to others, and to write explanations of their approaches.

- *Rich opportunities:* The tasks should let children solve problems via a variety of creative strategies; demonstrate talents (artistic, spatial, verbal) beyond those normally associated with numerical mathematics; invent mathematics that (to them) is new; recognize opportunities to use and apply mathematics; and show what they can do (as opposed to what they cannot do).

- *Improved instruction:* The tasks should have the potential for influencing instruction positively, both in content and in pedagogy. Teachers who use these tasks should become better teachers as a result of the experience; children who participate should emerge with increased self-confidence and heightened expectations for future mathematics courses.

The Caveats

These tasks are *prototypes*, not tasks ready for immediate administration to fourth-grade students. They are intended to illustrate possible directions for new assessment instruments, not to be an example of a real assessment. Certainly they should be viewed as work in progress, not as fully completed blueprints.

Criteria related to cost, efficiency, and immediate feasibility were deliberately not imposed on the work of the writing group. These are important considerations for implementation, but not for this volume. The MSEB goal for *Measuring Up* is to promote long-term change, not to write assessment material for current courses.

As assessment instruments, these prototypes are intended for children who have had the full benefit of a *Standards*-caliber mathematical education in kindergarten through fourth grade. Hence the tasks as presented here will be more appropriate, generally speaking, for students of some time in the future. From the perspective that has historically dominated U.S. testing, these prototypes illustrate directions for tomorrow, rather than tasks for immediate practical use. From a perspective more common in Europe — where tests, appropriately publicized in advance, set targets for teaching and learning — these prototypes do serve the immediate purpose of defining appropriate goals for fourth-grade instruction.

What we are not trying to do

Moreover, the prototypes, as a set, are not intended to illustrate a single assessment that treats all of the mathematics important at the fourth-grade level. Much that is important in the curriculum is not covered adequately in the particular examples chosen for this volume. Nevertheless, to expand our view of appropriate mathematics goals for the primary grades, these tasks provide more opportunities for children to demonstrate their ideas in areas often missing from the curriculum (e.g., data, geometry) than in areas already well entrenched (arithmetic). The imbalance in these examples reflects our desire to illustrate the new, not an effort to reshape the curriculum to fit this particular set of examples.

These prototypes, which are tasks to be done in time span ranging from one to three class periods, represent only one c many important forms of assessment. Other forms of assess ment are essential for a balanced program, including *projec* (extended pieces of mathematical investigation designed t take a substantial block of time), *portfolios* (structured collec tions of student work gathered over a long time period), an *tests* (time-limited responses to shorter tasks). Some of the re erences at the end of this volume (e.g., Pandey [1991] Stenmark [1989]) describe these alternative approaches.

The Audience

Many readers of *Measuring Up* will be persons who are pr fessionally concerned with mathematics education, particular developers of tests and other assessment instruments. For suc people, both those who work withi commercial test development con panies as well as those in educatio al settings at the state or local levels, *Measuring Up* should stin ulate development of new approaches to assessment that refle the broad goals of the nation's standards for mathematics ed cation.

Whom we are trying to reach

If mandated assessments evolve to resemble more close the ones suggested in this book, it is clear that differe approaches to instruction and testing will be needed. Hen school administrators and educational policy makers will al be affected by the changes implicit in these prototypes. Tl tasks will convey to the audience of policy makers and educ tion leaders what mathematics educators mean by assessme reform.

A third audience for *Measuring Up* consists of classroo teachers, and not just those at the fourth-grade level. It is or natural that many practicing elementary school teachers may fi some of these tasks to be somewhat daunting, especially if th students have not had the mathematical preparation that the tas assume. Teachers should look at the prototypes not as curre expectations, but rather as goals to aim for. The prototypes c be viewed both as examples of what tomorrow's assessment

mathematics might be like, and as examples of what today's goals for instruction should be like. In the meantime, teachers can use them as ideas for instructional activities for today. (A list of resources for teachers including the names and addresses of contacts in each state appears at the end of the volume.)

Another audience is the community of university-based educators who are responsible for the pre-service education of prospective teachers. They will find *Measuring Up* to be a source of ideas to use today for connecting the tenets of the mathematics education reform movement to classroom practice.

Finally, of course, the ultimate audience for these new assessment tasks and the ideas that underlie them is the elementary school children for whom the tasks were designed. The tasks provide good examples of challenging mathematical problems and situations that effective teachers can use even now as part of their instructional strategies. Today's children can begin to see the challenge in authentic mathematical problems even before tomorrow's tests give them an opportunity to demonstrate their accomplishments.

The Prototypes

Measuring Up contains thirteen assessment prototypes that exemplify changes called for in the *Standards*. In some cases the particular settings or contexts for the tasks are original, while in other cases some aspect of the task has appeared in another form previously.

The tasks in *Measuring Up* are intended for a largely hypothetical audience: fourth-grade children who have had a K-4 mathematics experience fully consonant with the NCTM *Standards*. Unfortunately, very few U.S. fourth graders have had the benefit of such an education. This is, of course, the whole point of the reform effort. One would not expect many of today's fourth graders to do very well on these tasks. Nonetheless the aim was to keep the tasks accessible to most of today's fourth graders; they should at least be able to understand what the tasks are about and become engaged in working on them.

What we have accomplished

Too often test questions and assessment tasks are presented solely in written form, which may be a burden for poor readers and for children whose first language is not English. Such children might not be able to respond to the tasks in a way that shows their true level of mathematical knowledge or skills. Many alternative presentations are possible: videotaped introduction; teacher-taught introduction; computer-based presentation; and presentation using manipulative materials. The prototypes illustrate each of these alternative modes of presentation, and two of the tasks are written in Spanish as well as in English.

Notwithstanding the possible variety in presentation, the prototypes in *Measuring Up* adhere to a certain uniformity of structure. Most are organized as a sequence of questions, often of increasing difficulty. On the one hand, this provides a structure around which the child's problem solving must be organized. On the other hand, this sequence of questions may suggest that the problem-poser, rather than the problem-solver, is in charge of the problem-solving process. Although other forms of organization are certainly possible, these prototypes provide sufficient imposed structure to help the mathematically less sophisticated student get started and show what he or she can do, while allowing plenty of open ended space at the top to challenge the more advanced student. Even though the questions within a task often grow in difficulty, many of the tasks involve problem solving, reasoning, and communication right from the beginning. These are important aspects of mathematics for all children.

Just as the tasks are presented in several formats, so they are also designed to give children a chance to respond in a variety of modes — perhaps by constructing an [redacted] or by creating a pattern on a computer screen. One imp[redacted] response mod

that is not specifically included in these prototypes is that of the child talking individually to a teacher, explaining his or her solutions orally rather than in written form. Pilot testing of the tasks has shown that children who have not had considerable experience in organizing their thoughts on paper find it much easier to tell someone else what they are doing than it is to record it in writing. Teachers who use tasks like the ones in this collection for their own informal assessment of how children are progressing mathematically will want to supplement written responses with spoken ones. In fact, asking a child to explain a solution in two forms — spoken and written — can help the child to sharpen and deepen both responses.

These prototypes can be used either for informal classroom-based assessment by an individual teacher, or for more formal external assessment, although certain modifications may be necessary to make the tasks suitable for a given purpose. All of the prototypes call for responses that go well beyond simple numerical answers, and most require the student to explain underlying patterns, relationships, or reasoning. As a result, the same activities can be useful to an individual teacher as she or he tries to discern more deeply how students are progressing mathematically, and to a district to discern the effectiveness of its instruction.

As the NCTM *Standards* urge, assessment should be embedded in instruction, so that most children would not recognize the assessment activity as a "test." Even when certain tasks are used as part of a formal, external assessment, there should be some kind of instructional follow-up. As a routine part of classroom discourse, interesting problems should be revisited, extended, and generalized, whatever their original sources.

Increasingly, educators are recognizing the value of having children work together in groups. Certainly group work exemplifies the NCTM's goal of stressing mathematics as a means of communication. Some of the tasks in *Measuring Up* are designed to be carried out in small groups, while in other cases, small groups are certainly a reasonable option. Continuing experimentation will be required to determine how the children can best be grouped for assessment tasks like these, and how to weigh individual vs. group work in performance evaluation.

In several cases the problems suggested here for fourth grade could also be asked in the eighth or even the twelfth grade, although naturally the expected sophistication and completeness of the responses would be very different. If a mathematical task is genuinely interesting and worthwhile for fourth graders, there is no reason why it should not be worthwhile for older children, or even for adults.

The Tryouts

Each prototype was tested on several score fourth-grade students in a number of different locales. These "tryouts" were not designed to be either representative or comprehensive, but to aid in improving the tasks. This they did, but they did much more as well. By observing how students react to the prototypes, we learned much about the gulf that separates current students from the goals of the *Standards*. We also learned that we are novices on how these new forms of assessment will work in the classroom.

What we learned from children

Three examples can illustrate the types of surprises that all teachers will encounter as they begin to explore and extend these prototypes:

- In a few cases the tasks as originally presented seemed not to be sufficiently challenging. One example is the "Lightning" task in which a fairly large proportion of the students could easily handle the map-reading requirements. So a question dealing with locating a lightning bolt that is a given distance from two observers was added.

- Sometimes a proposed task yielded no information of any interest at all. In "Bridges," there was originally a more open-ended question in which students were to create their own bridges. Nobody created anything that went even a little bit beyond the two-support, single-span examples. This may have been due to the wording of the question, to the backgrounds of the particular students, or to some other factor. This lack of inventiveness and perseverance is something worth pursuing since creativity i

an essential part of doing mathematics, for fourth graders as well as for everyone else. However, since the question produced virtually no information, it was dropped.

- One whole prototype was dropped entirely. It was a task on what is known as "Pick's Theorem" — which relates the area of a polygonal region on a geoboard to the number of nails on the boundary and in the interior of the region. The task was extremely open-ended and required extensive interaction between the teacher and individual students or small groups of students. Even if one assumed (as we do) that the teachers involved in the assessment are uniformly well versed in the subtleties of the underlying mathematics, there seemed to be no way of separating the effects of the teacher from the progress that individual students might make on the task. Perhaps such a task could be viewed as an assessment of the teacher-class unit, but in any case it seemed to be too problematic to include in this collection.

The Format

Each of the thirteen tasks is presented using the same outline. After the title, there is a suggested *time allotment,* which can vary from one to three class periods. This is followed by a suggested *student social organization,* although in many cases the task does not depend substantively on a particular form of grouping.

How we present the prototypes

Next comes the task itself. First there is a description of *assumed background.* In most cases this refers to specific aspects of the children's mathematical background, assuming — hypothetically, of course — that the children have had a K-4 education that fully meets the NCTM *Standards.* Second, there is a section on *presenting the task,* which details exactly what the teacher (or other assessor) should do. Finally, there is the *student assessment activity.* Very often this involves one or more sheets of paper on which students record their responses. (To reproduce these pages, which are scaled to the 7" x 10" page of this volume, the copying machine should be set to magnify them appropriately.)

The next major section is a *rationale* for the mathematics education community, which in many respects is the heart of *Measuring Up*. This is where comments on the content, style, or intent of the task appear (e.g., why the task was included), as well as more general messages about mathematics education that the task is intended to convey.

Following the main presentation of the rationale for the task, there are two subsections that provide further information. The first, task design considerations, discusses some of the details behind the task — why certain questions were phrased as they were, or why particular numbers were chosen. The second, variants and extensions, hints at other directions in which the task could be taken, for purposes either of instruction or further assessment. These subsections are far from exhaustive, for often the tasks could be starting points for weeks of instruction. One important message conveyed by this section is that these particular prototypes are in no way unique.

The next section describes a rough scoring system — what is called a *protorubric* — for the task. It is now widely recognized that an assessment task by itself means little without an indication of how children's responses would be scored. In other words, an important component of an assessment task is a scoring rubric that describes and orders a variety of answers that a child might typically give. For reasons discussed in the next section, the rubrics given here are necessarily tentative and incomplete — whence "protorubrics."

Finally, in some of the tasks there is a section containing *references* to relevant sources.

The Protorubrics

Although each task in this volume contains commentary about scoring based on student work, for a number of reasons we have not developed fully detailed scoring rubrics:

How might fourth graders do?

- The intended audience for these tasks are students who have had a mathematical educa[illegible]hat is different from what is commonly available in schools today.

Ideally, a scoring rubric should be based on the responses of many hundreds of children who are properly prepared for the tasks. While all of these tasks have been pilot tested with children, in most cases the testing has not been sufficient to provide a solid base for a complete scoring rubric.

- There is no universal agreement on how to structure scoring rubrics. Various groups who are currently active in creating alternative assessments in mathematics have used different styles and different levels of specificity (for example, four vs. six levels of gradation) for scoring rubrics.

- A complete analysis of scoring rubrics would require a foray into the thorny problem of judging individual performance in group settings. Although we do intend that these prototypes will encourage teachers to use group work, we have deliberately set aside the daunting task of codifying rubrics for assigning individual grades when students work in groups.

- There is continuing debate between proponents of "holistic" and "analytic" approaches. Does one look at every isolated component of a complex response, or should one make a general, overall, judgment of the child's response? While it is important to be fairly specific about what the task is intended to elicit and about what is to be valued in children's responses, there is no compelling evidence to favor one position over the other. The protorubrics given in this book can easily be adapted to different styles.

Moreover, protorubrics are in some ways analogous to standards: they express goals, ensure quality, and promote change in assessment. Hence, protorubrics by themselves may have a unique contribution to make to assessment reform, whether or not they ever are formalized into polished rubrics.

The protorubrics in *Measuring Up* are structured around three levels: high, medium, and low. Rather than try to define precisely what constitutes a "high" response, the protorubrics list only selected characteristics of a high response. We leave to others the

additional steps required to turn these outlines into fully detailed scoring rubrics and to refine the levels of response to each task.

One important purpose of creating a scoring rubric is to communicate to the *students* exactly what is expected of them. Embedded in our assumption that the students have had an exemplary mathematics education is an implication that appropriate standards have already been communicated to the students. Thus, for example, when a protorubric mentions a "clear explanation" or an "appropriate drawing," it is assumed that the children and the assessor share a common understanding of what these terms mean.

Another purpose to is to help the *teacher* interpret students' responses by specifying or clarifying the mathematical essence of the task — which aspects of the task are critical mathematically and which are not. These clarifications will be improved as tasks such as these are tested with larger numbers of students, particularly with those who have studied in a *Standards*-based curriculum.

The Standards

Are we measuring the right things?

Since this entire project has been undertaken in a context of mathematics education reform, an important question that naturally arises is the extent to which these prototypes reflect the spirit of the NCTM's *Curriculum and Evaluation Standards for School Mathematics*. Figure 1 suggests how these particular tasks relate to the content that the *Standards* calls for in grades K-4.

Having constructed this figure, we must emphasize how potentially dangerous such tables can be because they promote a "check-off" approach that conflicts with a truly integrated view of mathematics. Each "x" within the body of the table is merely shorthand for a detailed account of how the particular task exemplifies, or illustrates, or even extends the ideas within that particular standard.

In some cases, the "x" means only that the idea is possibly, but not necessarily, involved in the task. For example, an "x"

appears in the intersection of the "Hog Game" task and Fractions and Decimals because, as the protorubric states, one effective approach to the question about competing strategies depends on calculating each player's expected score, and this will require work with fractions or decimals. Similarly, children might create fractions to knock down pins in the "Bowl-A-Fact" task, and fractions could arise as part of finding an average number of buttons per person in "How Many Buttons?" Indeed, any sufficiently rich mathematical problem will allow for a variety of different approaches, and so the mathematics actually used may vary from one student to another. (On the surface, this appears to pose yet more difficulties for grading and judgment since student responses may be entirely satisfactory even while ignoring the skills supposedly being examined. Taking the broader view, however, the aim in these prototypes is to assess mathematical power, not individual specific skills.)

It is clear from the chart that the tasks have been designed so that each of them touches several of the NCTM *Standards*. (Note in particular that every task involves the four all-pervasive standards of problem-solving, communication, reasoning, and connections.) Of course it is a deliberate goal of these particular tasks to emphasize that mathematics is a connected and coherent discipline. Assessment tasks designed to involve many areas within mathematics will promote the parallel idea that instructional activities should also cross boundaries between topics.

Figure 1 also shows how the tasks are arrayed with respect to some aspects of mathematics that go beyond the NCTM *Standards* for K-4. Two of these — discrete mathematics and algebra — appear as components of standards in higher grades. Drawing attention to them here is meant only to suggest that

igure 1 *NCTM Standards* and Prototypes of Fourth-Grade Assessment Tasks

K-4 Standards	Hog Game	Towers	Bears	Lightning	Taxman	How Many Buttons?	Quilt Designer	Point of View	Bowl-A-Fact	Hexarights	Bridges	Checkers Tournament	Mystery Graphs
1. Problem Solving	x	x	x	x	x	x	x	x	x	x	x	x	x
2. Mathematics as Communication	x	x	x	x	x	x	x	x	x	x	x	x	x
3. Mathematics as Reasoning	x	x	x	x	x	x	x	x	x	x	x	x	x
4. Mathematical Connections	x	x	x	x	x	x	x	x	x	x	x	x	x
5. Estimation	x		x	x		x		x		x	x		x
6. Number Sense and Numeration	x	x	x	x	x	x			x	x	x	x	x
7. Concepts of Whole Number Operations	x		x	x	x	x			x	x	x	x	
8. Whole Number Computation	x	x	x	x	x	x			x	x	x	x	
9. Geometry and Spatial Sense		x		x			x	x		x	x		
10. Measurement		x	x	x						x	x		x
[illegible]tistics and Probability	x	x	x		x	x						x	x
[illegible]ractions and Decimals	x			x		x			x				
13. Patterns and Relationships	x	x	x	x	x	x	x	x	x	x	x	x	x
14. Discrete Mathematics	x	x	x									x	
15. Algebra											x		
16. Proof	x	x			x					x			x

[illegible] NCTM Standards

certain ideas from algebra and discrete mathematics are indeed appropriate in the lower grades.

The third, proof, appears in several of the tasks. Even in fourth grade, children should be given opportunities to formulate simple but convincing arguments. Statements that begin like these:

> "Johann could not *possibly* have gotten a 6 because . . . ."
>
> "There are *exactly* eight different three-block towers that can be made from two colors because . . . ."
>
> "You *can't* make a hexaright with an area of 36 cm^2 and a perimeter of 24 cm because . . . ."

can be completed in ways that amount to informal proofs.

The Future

These prototypes reveal just one aspect of the unfolding picture of reform in mathematics assessment. Both the NCTM and the MSEB are currently engaged in further efforts to promote standards-based assessment. NCTM is preparing assessment standards to complement earlier standards for curriculum and instruction. MSEB, in a parallel effort, is examining measurement and policy issues involved in various forms of assessment. In addition, advances in mathematics assessment are being made in many states across the nation.

Where do we go from here?

The messages conveyed by the prototypes in *Measuring Up* are consonant with the national goal of standards-based educational reform. In no way, however, do these prototypes provide definitive answers to the very deep and difficult issues surrounding assessment in mathematics education. The goal of *Measuring Up* is more modest: to further reform by providing rich examples that can be discussed and debated, refined, and improved. Through these prototypes we can glimpse the future of assessment in America.

Goals for School Mathematics

The 1989 NCTM report *Curriculum and Evaluation Standards for School Mathematics* identifies five broad goals for students' study of mathematics:

- *To value mathematics.* Students must recognize the varied roles played by mathematics in society, from accounting and finance to scientific research, from public policy debates to market research and political polls. Students' experiences in school must bring them to believe that mathematics has value for them, so they will have the incentive to continue studying mathematics as long as they are in school.

- *To reason mathematically.* Mathematics is, above all else, a habit of mind that helps clarify complex situations. Students must learn to gather evidence, to make conjectures, to formulate models, to invent counterexamples, and to build sound arguments. In so doing, they will develop the informed skepticism and sharp insight for which the mathematical perspective is so valued by society.

- *To communicate mathematics.* Learning to read, to write, and to speak about mathematical topics is essential not only as an objective in itself — in order that knowledge learned can be effectively used — but also as a strategy for understanding. There are no better ways to learn mathematics than by working in groups, teaching mathematics to each other, arguing about strategies, and expressing arguments carefully in written form.

- *To solve problems.* Industry expects school graduates to be able to use a wide variety of mathematical methods to solve problems. Students must, therefore, experience a wide variety of problems that vary in context, in length, in difficulty, and in method. They must learn to recast vague problems in a form amenable to analysis; to select appropriate strategies for solving problems; to recognize and formulate several solutions when that is appropriate; and to work with others in reaching consensus on solutions that are effective as well as logical.

- *To develop confidence.* The ability of individuals to cope with the mathematical demands of everyday life — as employees, as parents, and as citizens — depends on the attitudes toward mathematics conveyed by school experiences. One of the paradoxes of our age is the spectacle of parents who recognize the importance of mathematics yet boast of their own mathematical incompetence. Mathematics can neither be learned nor used unless it is supported by self-confidence built on success.

Report Offers Glimpse of Mathematics Assessment of the Future

By Robert Rothman

WASHINGTON—Providing a glimpse of the future of mathematics assessment, the Mathematical Sciences Education Board last week released a report describing 13 "prototype" math-assessment tasks for 4th graders.

While not exhaustive, the tasks are aimed at showing different ways of measuring the type of learning called for by the standards for the field developed by the National Council of Teachers of Mathematics, according to Nancy S. Cole, the report's author.

"The N.C.T.M. standards, and other math-reform efforts, are directed toward new kinds of learning for students in mathematics," said Ms. Cole, the executive vice president of the Educational Testing Service. "For those new goals, we have to have very different kinds of assessments."

Although several schools have begun to revamp their assessment systems to match the reforms, Ms. Cole noted, a recent study funded by the National Science Foundation found that most tests do not reflect those goals.

Group describes 13 'prototype' tasks keyed to N.C.T.M. standards.

In contrast to traditional tests, the prototype tasks included in the new report ask students to solve problems using reasoning skills, and to communicate their answers in writing or some other form, rather than simply to choose the correct answer.

Taking from one to three class periods to complete, they include a range of activities that involve both pencil-and-paper work and manipulative materials. One involves the use of a computer-graphics program.

In addition to the tasks themselves, the report also includes background activities to prepare students for the tasks, rationales that indicate the skills and knowledge they attempt to tap, and characteristics of high, medium, and low performance.

But while teachers can use the tasks in classrooms—in fact, they were all tested in elementary schools in four states—they are intended primarily for a "hypothetical audience": students who have been through a revamped math program. For current students, it says, the tasks may be "daunting."

"Teachers should look at the prototypes," the report states, "not as current expectations, but rather as goals to aim for."

Not a Ready-Made Assessment

The report issued last week, "Measuring Up," is a direct follow-up to a national mathematics assessment "summit meeting" held in Washington in April 1991.

At that meeting, the report notes, Gov. Roy Romer of Colorado, then the chairman of the National Education Goals Panel, "challenged the mathematical community to show the nation what mathematics educators mean by mathematical power and what new and more demanding standards will mean for our young people."

In response, the N.S.F.E.B., an arm of the National Academy of Sciences, agreed to develop a set of prototypical assessment tasks, and convened a writing group consisting of math educators from universities, schools, and research centers.

The report cautions that the document—which it calls a "work in progress"—is not intended to present an assessment that could be administered immediately.

For one thing, it notes, the panel was not asked to consider issues of cost, efficiency, and feasibility.

In addition, the report, by focusing on tasks, does not include other forms of assessment, such as projects and portfolios, and does not represent the full range of topics that 4th graders should know and be able to do, it states.

"Much that is important in the curriculum is not covered adequately in the particular examples chosen for this volume," the report states.

Variety of Formats

In presenting the tasks, the report notes, the authors set out to show that assessments can be introduced in a variety of formats: by videotapes or by teachers, with manipulative materials or computers. The report also includes two tasks that are presented in Spanish as well as in English.

"Too often," it says, "test questions and assessment tasks are presented solely in written form, which may be a burden for poor readers and for children whose first language is not English."

It also notes that the tasks allow students to respond in different ways, including constructing an object and creating a pattern on a computer screen.

In keeping with the N.C.T.M. standards, the tasks differ sharply from traditional math tests.

Their content, for example, incorporates a variety of mathematics, particularly topics—such as statistics, geometry, and probability—that are seldom emphasized at the K-4 level, the report notes.

One task, for example—which asks students to analyze a graph network related to a checkers tournament—was included partly to show that some mathematics does not involve computation, the centerpiece of the traditional elementary curriculum.

Allowing Students To Decide

In addition, all of the tasks offer students the ability to decide how to solve problems and to explain how they went about solving them.

In fact, the scoring rubrics indicate that students who provide well-reasoned explanations for their responses earn higher scores than those who simply provide the correct answer.

Moreover, in some cases, the report points out, there is no "right" answer.

"A question is not very good if the right answer is a single number," Ms. Cole said. "There are better answers and less-good answers."

Several of the tasks also allow students to decide how they will answer questions.

One task, which asks students to draw a new geometric figure called a "hexaright," deliberately leaves too little space to answer the question.

"The purpose of this is to force the child to decide what kind of paper to use," the report states. "Centimeter graph paper will be helpful to some students and a distraction for others."

In another task, which asks students to generate equations to solve an arithmetic game called "bowl-a-fact," the questions leave unstated whether a calculator should be used. In the N.C.T.M. standards, the use of calculators is an important skill.

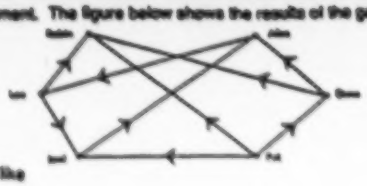
But while calculators should be available to students, the report notes, "it will soon become clear to them that calculators have very limited value in this situation. In fact, using a calculator to generate an equation to knock over pins is an extremely inefficient way to approach the task."

"Thus," it says, "one reason for using these kinds of tasks is to sharpen students' perception of when calculators are useful and when they are not."

Copies of the report, "Measuring Up: Prototypes for Mathematics Assessment," are available for $10.95 each, prepaid, plus $4 for shipping, from the National Academy Press, 2101 Constitution Ave., N.W., Washington, D.C. 20418.

Reprinted with permission from EDUCATION WEEK, Volume 12, Number 14, December 9, 1992.

Six children are in a checkers tournament. The figure below shows the results of the games played so far.



(Remember, in the picture, an arrow like

means that Jose won his game against Alex. The arrow always points from the winner to the loser.)

1. Who won the game between Pat and Robin? _____
2. Which children has Les already played against? _____
3. Which of those games did Lee win? _____
4. How many games have been played by the children so far? _____
   Explain how you know.

5. Make a table showing the current standings of the six children. Put the player who has the most games in first place, at the top. If two players are tied, they can be listed in either order.

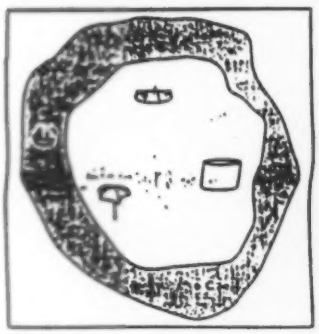
Name	Wins	Loses
1.		
2.		
3.		
4.		
5.		
6.		

6. The tournament will be over when everybody has played everybody else exactly once. How many more games need to be played to finish the tournament? _____
   Explain your answer.

7. Dana and Lee have not played yet. Who do you think will win when they play? _____
   Explain why you think so.

James has rented a rowboat to row in the pond around the playground. On the playground there are three pieces of equipment:

a play fort: an umbrella: and a merry-go-round:



Look at the map. Find James in his boat. Imagine how the playground looks to James. From where James was, the playground looked like this:

The merry-go-round was on the left, the umbrella was on the right, and the fort was in the middle.

1. James rowed in the pond for a while. When he looked at the playground a second time, this is what it looked like:

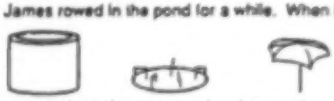


Figure out where James was when he saw the playground this way. Draw a dot there and label it A.

2. Then James rowed some more. He came to another spot, and stopped. When he looked at the playground from this spot, it looked different. From here, the playground looked like this:

Where was James now? Please draw another dot on the map to show where James was and label it B.

3. James really enjoyed rowing, so he rowed some more and explored some more. After a while he was at the spot marked C.

Please draw the fort, the umbrella, and the merry-go-round in the space below, to show the way the playground looked to James when he looked at it from the spot marked C.

4. It was a warm, pleasant day, and James thought that maybe he would keep on rowing all afternoon. He rowed to a new spot on the pond. When he looked toward the playground, he was suddenly surprised. He could no longer see the umbrella! All he could see was the fort on the left, and the merry-go-round on the right.

Please draw a dot on the map to show where James was when he discovered that he could not see the umbrella. Label it D.

5. Do you think that James could row to a spot on the pond where he sees the fort on his left and the umbrella on his right, but he can't see the merry-go-round? _____ Please explain how you know.

SOURCE: Mathematical Sciences Education Board BEST COPY AVAILABLE

THE POWER OF THINKING MATHEMATICS

Reprinted with permission from AMERICAN EDUCATOR, Winter 1992.

BY ALICE J. GILL AND LOVELY H. BILLUPS

IT WAS 2:30 on the afternoon of February 14, 1990. In an inner-city classroom in Rochester, New York, Chapter 1 Basic Skills teacher Marcy Miller stepped into the second-grade classroom prepared to conduct her twice-weekly math lesson. She approached the homeroom teacher.

"Mrs. Jones, I know you're having a Valentine's Day party this afternoon. Tell me how much time I have for a math lesson."

"Miss Miller," came the frustrated reply, "these children were so disorderly this morning, I don't think they're going to have a party. You can have the rest of the afternoon." A young girl immediately raised her hand.

"Does that mean we'll have math longer today?"

With conviction in her voice that bespoke a punishment well delivered, Mrs. Jones almost beamed as she affirmed, "That's right!" whereupon, the class began to cheer and applaud.

Seven-year-olds happily forsaking a Valentine's Day party for the opportunity to do more math is not a picture that typifies our classrooms. In fact, report after report casts American youngsters—and adults—as being uncomfortable with mathematics, indeed, often expressing an intense dislike for the subject.[1] Math, they feel, is best left to engineers, scientists, and a small elite group endowed at birth with a talent for the arcane world of numbers.

Alice J. Gill is an assistant director of the educational issues department of the American Federation of Teachers. As a third-grade teacher in Cleveland, Ohio, she was one of the developers of Thinking Math. Lovely H. Billups is director of field services of the AFT's educational issues department and coordinator of its Educational Research and Dissemination program.

It's not surprising that large numbers of Americans have an aversion to math. Most people dislike activities they're not good at, and on that score the figures are stunning. The inadequate achievement of U.S. students in mathematics has been chronicled in headlines, papers, books, and conferences. In a 1991 study conducted by the Educational Testing Service, American nine-year-olds ranked twelfth out of fourteen industrialized nations in math, ahead of only Slovenia and Portugal; American thirteen-year-olds ranked sixteenth out of twenty. On the 1988 International Association for the Evaluation of Educational Achievement, for ten-year-olds, 38 percent of American schools scored below the *lowest*-scoring school in Japan. On the 1991 National Assessment of Educational Progress, fewer than half of the twelfth graders demonstrated a consistent grasp of decimals, percents, fractions, and simple algebra.

Doing something about the American dilemma in math can at times feel like trying to move a mountain; it has been there a long time. Yet, though still rare, the real-life scene described in the opening of this article is becoming more common as efforts proceed to transform America's classrooms into places where children not only look forward to math but master it well enough to successfully compete on a global level.

WE GO NEXT to a classroom in Lake County, Florida, where students are thinking about a class picnic. They know that twenty-four students and seven adults will be present. One of their tasks is to decide how many six-packs of soda to buy so each person can have a can of soda. In small table groups, they move counters representing the cans and finally determine that they need to purchase half a dozen six-packs. The problem isn't extraordinary. It's a two-step problem that requires stu-

dents to make a substantive decision about how the remainder in a division problem should affect the answer. This is one type of problem with which American seventh-grade students have had difficulty on National Assessment of Educational Progress (NAEP) tests[2]. What is unusual in this instance is that the problem was successfully solved by these Lake County first graders.

The lesson was a demonstration of Thinking Mathematics, a research-based approach to teaching mathematics that grew out of a collaboration between the American Federation of Teachers (AFT) and the Learning Research and Development Center (LRDC) of the University of Pittsburgh. The philosophy of Thinking Mathematics, which is consistent with the standards adopted by the National Council of Teachers of Mathematics, is that we can produce not only students who are capable mathematicians but also a populace that appreciates the place of mathematics in their own lives and that is no longer "mathophobic."

In Thinking Math, value is placed on thinking, reasoning, communicating mathematically, focusing on relationships, using what is known to find the unknown. Its content and sequence is shaped by the idea that, right from the start of their mathematical education, children can and should be engaged in the discussion, analysis, and solution of mathematical problems; they need not wait until, step by step, they have mastered a strict hierarchy of basic skills. It draws heavily from the cognitive apprenticeship model, which emphasizes the importance of working on authentic, real-life tasks and of exposing students to the reasoning and strategies that experts employ when they acquire knowledge or put it to work to solve such tasks.

This approach is in sharp contrast to the milieu out of which teachers have come, which emphasized memorization and rote procedures. The majority of elementary teachers have scant background in mathematics. Because it has not been expected of them, few have taken anything beyond high school math, except for a college course in elementary math methods. The common image of the proper way to teach math, held by both teachers and the public, is based on their own memorization and formula-driven school experiences. On the elementary level, the core of activity in math classes centers on what are called "the basics." This consists of memorizing basic facts and rules and performing page after page of computation. Probably 90 percent of the math activities in elementary classrooms have been traditionally conducted to facilitate this memorization and calculation.

In addition, studies indicate that math instruction in the United States is repetitious and poorly organized. Teachers spend weeks and sometimes months each year repeating the content of previous years. In many instances the students appear to have "forgotten," when in fact, they may never have really learned the concepts.

We now know with certainty that this traditional approach to math education will not produce thinkers, interpreters, and users of information.

WHEN CHILDREN come to school, excited and curious, they already have ways of thinking about quantities and numbers. What usually happens is that they encounter a teacher who begins to model and, with the best of intentions, demand that the children mirror "the correct way" to think about and manipulate numbers. If this "correct way" is the child's way, the child becomes a star performer; if not, the child becomes confused, loses confidence in himself, and begins to form a disastrous opinion about his ability to do math. This is not to say that everything a child thinks is true or valid. Yet children play counting games, share, and make purchases at the store before they come to school. They think about numerosity and measurement in the world around them, choosing the biggest piece of pie, the bag with the most cookies, arguing about who's taller than whom. Using the perspective and knowledge that children bring to school is the first of Ten Principles that constitute the Thinking Mathematics approach to teaching mathematics. Although the phrase "start from where the children are" has echoed through the halls of schools for decades, it has referred to where they are in relation to our curriculum and not to how children think or learn.

Right from the start of their mathematical education, children can and should be engaged in the discussion, analysis, and solution of mathematical problems; they need not wait until, step by step, they have mastered a strict hierarchy of basic skills.

The Ten Principles of Thinking Mathematics are:

1. Build from intuitive knowledge.
2. Establish a strong number sense through counting, estimation, use of benchmarks, mental computation skills, and understanding the effects of operations.
3. Base instruction on situational story problems.
4. Use manipulatives and other representations to represent the problem situation; then link concrete and symbolic representations.
5. Require students to describe and justify their mathematical thinking.
6. Accept multiple correct solutions and, in some cases, more than one correct answer.
7. Use a variety of teaching strategies.
8. Balance conceptual and procedural knowledge.
9. Use ongoing, new assessments to guide instruction.
10. Adjust the curriculum timeline.

In addition, Thinking Math (TM) recommends that teachers focus on depth instead of quantity, that math classes be used to look at a few problems from many angles rather than to work many problems the same way. This belief, which grew out of the research findings, is similar to the practices of Japanese and Chinese classrooms where the goal is lasting conceptual understanding and where it is not untypical to devote an entire class period to one or two problems.*

The picnic soda scenario described earlier was one part of the exploration that day. In addition to attending to children's intuitions, the lesson visibly incorporated two other principles of TM. First, the students had physical objects (manipulatives) with which they could model and, thereby, strategize about, a familiar situation. Thus, numbers, quantities, and operations had meaning beyond paper. Secondly, the students were grappling with a problem that had meaning to them. In Thinking Math, teachers are urged to write or change problems to reflect their own classroom, school, or city and to incorporate activities about which their students express interest. This not only increases motivation but also shows that math is, indeed, connected to the real world.

As a classroom is opened to students' thinking, there begin to emerge multiple intuitive and inventive ways of solving problems. The encouragement and support of this process is a crucial part of the teaching and learning in TM classrooms. Again, there is a strong parallel to real-life situations. Woe to any commercial enterprise that sees only one way to solve a problem! If that one solution becomes stymied, the business may fold unless an inventive mind finds another way to come at the problem. Both students and teachers come to appreciate this principle. Nine-year-old Brandon, addressing the Anderson, Indiana, school corporation one night, observed, "In Thinking Math, there are a lot of ways of doing a problem so we can choose the *best* way."

In Albuquerque, New Mexico, another nine-year-old who had transferred to the public schools from a private one the year Thinking Mathematics was introduced, shared how he felt about the program. "My other school thought they were real good, but this one is better. They only taught us one way to do a problem, but we learned a lot of ways here." And still another student's voice: "What do I think about Thinking Math? It helps you think a lot better. Not only letting kids do it different ways but letting kids do it their own way . . . you can think with your head and do math the way you need to do it to solve problems and stuff like that."

These different strategies be dormant and unvalidated, however, unless they are shared within a community of learners. Thus, requiring students to explain and justify their strategies is another important Thinking Mathematics principle. This public discussion about mathematics (mathematical discourse) adds power to students and teachers in three ways. It requires the speaker to clarify his own thinking. It allows other students to get additional perspectives on a problem. It gives the teacher information about the level of understanding a student has, information that is not evident when a teacher looks only at computation.

A second-grade class in Lake County, Florida, had worked through a problem that centered on a typical elementary school Valentine's Day occurrence. Children bring in cookies and they are shared by the class. One of the goals of the lesson was to demonstrate how number sense can be used to simplify the solution in a way frequently used by adults but rarely taught in school. When adults buy two items that cost $1.98 each, they often calculate two times two dollars and subtract four cents rather than go through the regrouping algorithm. The problem the second-grade students were solving required them at one point to add 36 and 29, the numbers of cookies brought in by two students. A stuffed replica of Curious George (whom the teacher often used as a friendly voice to introduce other ideas into the discussion in a nonintimidating way) suggested that if you could add 30 instead of 29, it would be a really simple calculation, which he could do in his head.

"36 + 30 is 66."

"Could we do that?" asked the teacher. Not two seconds passed before young Jessica raised her hand and said, "Yes. But you have to take one away from that because you added one too many. So you'd have 65 cookies."

The role of the teacher in these exchanges is extremely important. Encouraging students to do their own thinking and to find many ways to solve problems is not the same as telling students to "go discover" while the teacher stands back and watches. The teacher becomes planner extraordinaire, framer of the circumstances that will enable students to find their way. Generally, the solution strategies are those that are brought forward by students as they think and use the materials provided. In this instance, the teacher interjected "another way" to think about the problem. For the process to avoid turning the classroom into a venue where the teacher is seen as custodian of "the right way," the teacher must know which concepts or strategies students will probably *not* find intuitively and make a judgment about when and how to introduce those ideas.

Finally, the teacher must be capable of moving in directions suggested by student conversation and of focusing that conversation on mathematical ideas, targeted and untargeted. To do this successfully, the teacher must develop a good sense of the mathematical territory and of students' conceptual understanding. The student's explanation must be clear enough for other students to follow the thought. In instances where it is not clear, or

where the teacher believes some students need a concrete model, the teacher appropriately raises questions or asks the solver to concretely demonstrate what has been said.

Leading mathematical discussions is a complex task. Students long accustomed to traditional math classes may be reluctant to express their ideas at first. The quality of discussion improves over time as students become aware that their ideas are valued, as they learn to express themselves, and as teachers, too, learn from actual conversations. Teachers must become knowledgeable about what actually demonstrates understanding. To return again to the 36+29 Valentine cookies problem, an explanation that basically does nothing more than repeat a formula (e.g., 6+9=15, put down the 5, carry the 1; 1+3+2=6; 65) does not show the same evidence of number sense that investing quantities with their proper meaning does (30+20 is 50; 6+9 is 15; 50+15 is 65).

One young man in Cleveland, near the end of the first year of Thinking Math, turned in the following as a solution to a subtraction problem.

541 - 268
500 - 200 = 300
41 - 68 = -27
300 -27 = 273

This youngster had not been taught this method. But he had been in an atmosphere that clearly valued exploring number territory and finding different ways. He would undoubtedly have been a good "traditional" math student. He became, for his short time in Thinking Mathematics, a marvelous inventor who saw how things fit together and what was happening with the quantities in this operation. He knew 41 minus 68 was minus 27, he said, because he took away 41 and then he still needed "fifty-one, sixty-one (*he had raised two fingers*) and 7 more for 68. So that's twenty (*holding up the two fingers, each of which stood for 10*) seven." This student's explanation encompassed a way of thinking about 41 minus 68 that his teacher would not have thought of.

WITH ALL the promise of the current mathematics scene, there still exists a good deal of skepticism. Parents worry about whether their children will learn basic facts and be able to compute, the centerpieces of their own math education. Those whose children are now grown worry about repeating the disastrous path of "new math" from the sixties. It is, therefore, important to clarify the relation of Thinking Mathematics to both of these concerns.

First, the concern about learning basic math computation skills. One of the Ten Principles of TM is to provide a balance between conceptual and procedural learning. This means that, where formerly it was acceptable if students were able to rotely compute even if they didn't understand what they were doing, this is no longer considered exemplary performance. Even the "good" students who went through "drill-or-kill" tactics are unable to compete with their peers worldwide. One of the deficiencies of these students is their inability to reason and solve complex problems.

Teaching for conceptual understanding takes longer than rote learning. It does, however, make it possible for students to also acquire basic knowledge without an overemphasis on unmotivating, unconnected drill and the calculation of bedsheet-length sets of problems. TM deemphasizes the amount of time spent on learning basic facts, believing that students will learn them as they work with them. Walk down the halls of a school where teachers are doing what they have always done and you will find a belief that "until students know those basics" it is a waste of time to turn their attention to something as complicated as word problems. "Thought problems" at the ends of textbook chapters are generally skipped because "there simply isn't time available if the students are to master their facts." Visit a TM class and you will find students engaged in solving problems about familiar situations; some know the facts, others are still learning them, but all can successfully solve interesting problems. You may find some periods of short (5 to 7 minutes) drill on facts. If you do, the drill will often have a thinking

BRANDON'S TEN GOOD REASONS TO LIKE THINKING MATHEMATICS

HI! I'M Brandon Sokol. I come from the family class at Robinson School. I am going to give ten reasons why I like Thinking Math better than traditional math.

1. Thinking Math eliminates the need to memorize a problem cause we don't learn to use paper, we learn to use our heads.
2. In Thinking Math we can make up and solve our own sichawashunol story problem.
3. We are also able to use manipulatives to help solve the problem. In Thinking Math, a manipulative is a small block or tool to help count with. This (displaying a base ten rod) is ten. This is a hundred. This is one. And this is a thousand.
4. We can also use decomposition to break down a problem. Decomposition is breaking down a number, for instance, 378 would be 300+70+8. That would be easier to add to another number.
5. In Thinking Math there are a lot of ways of doing a problem so you can choose the Best way.
6. In Thinking Math we learn to do our math in our head so when we go shopping we can add numbers fast.
7. We discuss a problem to make sure everyone knows the steps to the problem to complete the problem so that the next aren't so hard.
8. In Thinking Math we use our own record book of our knowledge not someone else's.
9. We only need to do 3 or 4 problems during math.
10. With Thinking Math a group of children can all do one problem together.

—BRANDON SOKOL
Age 9
Addressing the Anderson, Indiana, School Corporation Board of Trustees

The children are so excited! They're always asking to do math. I've never had that happen before.

schema behind it.

7+8
5+6
8+9

In this exercise, students learn thinking strategy for "near doubles." Doubles are learned fairly quickly. If students know the doubles (e.g., 7+7) they can project an answer that is one more than the double. Or, the drill may promote a mental math skill that will be useful in computation.

17+10
38+10
52+10
69+10

There is evidence that students who spend less time on drill and practice do not suffer when it comes to "the basics." In fact, the Cognitively Guided Instruction project, an approach similar to Thinking Math, found that students in classrooms with a problem-solving focus actually outscored their peers. Additionally, a study by SRI International found:

> By comparison with conventional practices, instruction that emphasizes meaning and understanding is more effective at inculcating advanced skills, is *at least as effective at teaching basic skills*, and engages children more extensively in academic learning (emphasis added).[8]

Even the results of standardized tests—which are not geared to measure the kinds of conceptual understandings Thinking Math aims for—put TM students above their non-TM counterparts. Researchers from the University of Pittsburgh compiled data from student scores on standardized tests for the group of TM pilot classes in 1990-91. The report[9] of this data stated:

> Standardized achievement test scores indicate that project students did at least as well, if not better, than their non-project peers on both Computation and Concepts and Applications subsections.

Class scores were averaged for all students in grades one through five for TM classes and non-TM classes for each of the subsections of standardized achievement tests administered in the local districts;[10] the results are presented below.

Subsection	TM classes	Non-TM classes
Computation	64	59
Concepts and applications	66	50

In addition, the report continued:

> Notable improvements in student problem-solving abilities were indicated by results of the problem-solving test.[11] The percentage of correct answers on the post-test exceeded that to be expected from the item difficulty established by the test makers.
>
> . . . (T)here are multiple indications that student learning and attitudes were enhanced by their participation in the Thinking Mathematics program.

The spontaneous observations that Thinking Math teachers have made not only support the data regarding student achievement gains but also point to qualitative differences as well. One teacher in Brevard County, Florida, observed, "They are so excited! They're always asking to do math. I've never had that happen before." Another commented, "I'm now teaching my children how to think and understand math, not just how to do math."[12] Still another from a pilot class observed, "My students are being exposed to and coming up with skills I would not have thought possible." Another reflected that the students were showing greater appreciation for the use math would have in their lives.

THE SECOND concern about a new approach to teaching mathematics is the lingering aura of distaste from the "new math" of the sixties. There are several ways in which Thinking Math differs. One is that "new math" was imposed upon teachers by people perceived to be out of touch with classrooms. A result of the actual top-down strategy of this movement is that it did not

provide, in time or in substance, what teachers needed to understand the process and the math they were being asked to teach. Because teachers stood at the core of the development of Thinking Mathematics, they were able to identify these needs. Thinking Math also has provided that framework of research that professionals can use to inform what they ought to do and why. While some concepts that were part of "new math" surface again, they surface with a rationale and a professional development effort that allows teachers to make meaning of them. Another prime problem with "new math" was that parents were left out of the equation.

> In discussing the resistance encountered by "new math," the National Research Council notes, "When parents could not or did not understand the need for change or the reasons new curricular emphases (of that program) were chosen, resentment and anger resulted, and a conviction set in that if the 'old math' was good enough for them, it was good enough for their children." Parents have become enthusiastic about programs when they have been given the opportunity to ask questions and have them answered, are involved in their children's learning, and are assured that their children's computational skills and learning outcomes are not suffering."

Thinking Mathematics advises its teachers how important it is to communicate with parents as they begin the program, and, on the whole, they have had a very positive response.

Not only has the developmental process for Thinking Math been different from the one for "new math," there are also substantive differences. As it was implemented in classrooms, "new math" concepts were taught abstractly. Since this is not how children best learn math, the goal of developing a deeper understanding was not achieved. For example, "expanded notation" (e.g., writing 485 as 400+80+5) was often used in "new math." But the instruction was generally based on abstract, contextless figures.

In Thinking Mathematics also, students frequently record solutions using expanded notation. However, this flows from a construction of the meaning of those quantities that starts when students manipulate and group objects; compose, decompose, and recompose numbers many ways *as* they solve problems about familiar things and in ways they understand *before* they are introduced to efficient algorithms. So their understandings of both quantities and operations are richer and more lasting and they can successfully attack unfamiliar situations.

The "new math" also lacked a surrounding system of student dialogue about mathematics that helps children develop and clarify their thinking while allowing the teacher to better see the depth of their understanding and where they falter. Nor was it grounded in a philosophy that accepted and promoted multiple ways of solving problems.

• • •

CHANGE IS never easy. The history of education reform is cluttered with great ideas that never took root in the classroom. Countless manuals, materials, and manipulatives, once imbedded in such hope, are sitting somewhere collecting dust.

Thinking Mathematics is one program that has not suffered such a fate. It is functioning in thirty-eight cities, with a waiting list of dozens more. A 1992 survey of participating teachers found that they were making significant changes in their mathematics instruction. On average, they were using strategies recommended by Thinking Math three times a week, or 60 percent of the time."

To understand why Thinking Math appears to be taking hold where other reforms failed, we have to go back a few years to the start of the American Federation of Teachers' Educational Research and Dissemination program. ER&D began in 1988 with the goal of enhancing teachers' professional development by giving them access to an expanding knowledge base of classroom research and involving them with the educational researchers doing that research. As AFT president Albert Shanker noted at the time, teachers will be denied recognition as professionals until they can demonstrate that their actions and judgments are grounded in a solid professional knowledge base acquired through intensive and continuing study.

In the past, much educational research was packaged in ways that were remote from teachers' daily experiences. Most studies were abstract and jargon laden, so it's not surprising that teachers tended to find them of little use, if they read them at all. When mandated from above, as many "research-based" teaching programs were, they were met with suspicion from teachers and produced superficial change, at best. For their part, few researchers seemed to know or care about teachers' perceptions and were unable to relate to the realities of the classroom.

AFT's ER&D program sought to reverse those negative trends, to encourage teachers to value the information available from authentic research findings, and to expand their tools of practice. Originally funded by the National Institute of Education, ER&D is a long-range, peer-to-peer, union-sponsored strategy for professional improvement that encourages teachers to become users of research.

A KEY COMPONENT of the program is the development of "research translations" that highlight and interpret the most important research findings in practical ways that teachers can use; translations have been done on a wide array of topics, from cooperative small-group teaching to student motivation. The translations, based on a single work or a group of related works, eliminate ponderous statistics, interpret technical language, and focus on the practical applications of the research. As part of the process of developing ER&D materials, experienced teachers work collaboratively and intensively with researchers from universities and educational research laboratories.

But the translations are only the first step in the process of putting valuable research findings in teachers' hands. Through the ER&D network, teachers in local AFT affiliates receive training in the basic translations and in peer-teaching techniques. Once trained, these teachers in turn establish local programs and train other teachers, spreading the program's benefits to a growing number of classroom teachers. The local trainers meet period-

The bottom line of this endeavor is not books or manuals, but the deeper question of actually altering what goes on in the classroom.

cally with AFT national staff to review material, share their experiences, and learn about new research translations. A three-part philosophy guides the ER&D training: Sessions are non-threatening, non-judgmental, and voluntary.

Thinking Math provided the ER&D program's first move into specific subject matter; previous translations had dealt with more generic teaching skills. Started in 1987 through a National Science Foundation grant to the AFT and the Learning and Research Development Center at the University of Pittsburgh, Thinking Math built on the ER&D tradition of close collaboration between teachers and researchers. Lauren Resnick, LRDC's co-director and one of the first researchers to interact with ER&D teachers, describes the experience this way:

> The ER&D program enables teachers to probe, question, and interact with educational researchers about their findings. It is unique because it does not create a difference in rank or prestige between researchers and teachers. It stresses that researchers and teachers need each other—that educational effectiveness is not the exclusive province of either group.

Combining the clinical wisdom of teachers and the rich research background of the cognitive scientists, the collaboration has to date produced two *Thinking Math* volumes, covering counting, estimating, adding, subtracting, multiplying, and dividing. The joint effort has also resulted in the publication of a recent book, *Analysis of Arithmetic for Mathematics Teaching* (Lawrence Erlbaum), that discusses current research knowledge relevant to teaching math in grades one through eight.

THE BOTTOM line of this endeavor, however, is not books or manuals, but the deeper question of actually altering what goes on in the classroom. And here the evidence is cause for optimism. For example, as the survey mentioned above found, only 19 percent of the teachers said they had encouraged their students to solve problems in more than one way before becoming involved with the program; after their TM training, 71 percent were focusing on multiple strategies.

The effectiveness of the training depends in large part on the degree of cooperation between the local affiliate and the school district. The districts where the program appears to be flourishing have provided teachers release time not only for initial training, but also for regular follow-up sessions.

One of those districts is Anderson, Indiana, where the various partners have shared the cost of building a solid Thinking Math program over the past few years. The AFT has paid for training two local Thinking Math coordinators, and the district has picked up the tab for training seventeen other local teachers (including the cost of substitutes and release time once a month to allow teachers to reflect together on their experiences). The Anderson Federation of Teachers has also contributed significant amounts of money in the past three years to expand the program. By the end of this school year, every elementary teacher in Anderson will have had an introductory session on Thinking Math.

The experience of one Anderson teacher of learning-disabled students provides an insight into the power of the Thinking Math approach. The teacher convinced three of her colleagues that all their students could learn together if they taught math the TM way. With the teachers working as a team, the program has proved so successful that it is difficult for an outsider to distinguish the learning-disabled students from their classmates during math lessons.

In Albuquerque, New Mexico, teachers at one school have seen the power of Thinking Math in helping them break their professional isolation. The teachers arranged their master schedules to provide periods of time for teachers of the same grade to collaborate and talk about their math lessons. In the philosophy of the Japanese

(Continued on page 48)

teachers who carefully hone each lesson to perfection, they call this their "polishing time."

Rhode Island provides another promising model. There, the Rhode Island Federation of Teachers has brought a Thinking Math team to the state to work with teachers from several districts. As the local teachers wrestle with their own initial implementation of the program, they will be able to meet monthly throughout the year, which will provide a level of mutual support for the teacher leaders that had not been possible for other teams.

THESE STORIES illustrate the kind of commitment that Thinking Math requires. The program asks teachers to rethink their most basic beliefs and assumptions about teaching and learning mathematics. Such radical change cannot be brought about by one-shot "professional development" workshops or by plopping manuals into teachers' laps. There is no substitute for the collegial and research-based process that permeates the ER&D training; it provides a forum and support network for solving the problems that arise when teachers make substantive changes. To get past the inevitable bumps in the road that accompany change, there also needs to be a non-threatening atmosphere, sufficient training that continues after the initial training is done, and opportunities for regular interaction with colleagues.

Some teachers who are trained in Thinking Math have been troubled when they do not return home with a set of discrete and sequenced activities. But they come to realize that the program requires that they reconstruct their teaching, using their local curriculum, from their new knowledge and beliefs. More than 90 percent have been able not only to make this adjustment for themselves but also to successfully inspire, train, and pass on their ability to their peers.

Reforming the way mathematics is taught would be accomplishment enough for the program. But its effects are more sweeping. Thinking Math has convinced those intimately involved with it that the best route to genuine education reform is through a new look at content. When teachers passionately believe that new approaches are necessary and productive in their daily teaching (of math, in this case), they begin to see that changes must be made in the entire structure of schools to accommodate and support those new approaches. They begin to rethink how the school day should be organized; they come head to head with standard assessment practices and realize they need overhauling; they redesign staff development and consider new ways of organizing school staff. Whatever stands in their way gets close scrutiny, and what starts as Thinking Math often adds up to much more. □

REFERENCES

[1] Mathematical Science Education Board of the National Research Council (1989). *U.S. School Mathematics from an International Perspective: A Guide for Speakers.* Washington, D.C.: National Research Council.

[2] Kouba, V.L., Brown, C.A., Carpenter, T.P. Lindquist, M.M., Silver, E.A., & Swafford, J.O. (1988b). "Results of the Fourth NAEP Assessment of Mathematics: Numbers, Operations, and Word Problems." *Arithmetic Teacher, 35*(8), 14-19.

[3] Collins, A., Brown, J.S., and Holum, A. "Cognitive Apprenticeship: Making Thinking Visible." *American Educator,* Winter 1991. Washington, D.C.: American Federation of Teachers.

[4] Stigler, James W., Stevenson, Harold W. "How Asian Teachers Polish Each Lesson to Perfection," *American Educator,* Spring 1991, American Federation of Teachers, Washington, D.C.

[5] National Assessment of Educational Progress (1989). *A World of Differences: An International Assessment of Mathematics and Science.* Princeton, New Jersey: National Assessment of Educational Progress.

[6] Knapp, M.S., Shields, P.M., Turnbull, B.J. *Academic Challenge for the Children of Poverty: Summary Report.* Washington, D.C.: U.S. Department of Education, Office of Planning, Budget, and Evaluation, 1992.

[7] Hojnacki, S.K. & Grover, B.W. "Thinking Mathematics: What's in It for the Students?" Paper presented at AERA, San Francisco, 1992.

[8] Hojnacki and Grover, Op cit. The score for each class was the average national percentile score attained on the standardized test.

[9] Wood-Cobb Problem-Solving Test

[10] Osborne, M. Practicum report, 1992

[11] Bodenhausen et al. *Thinking Mathematics,* Volume 1: Foundations, p. 5"

[12] Gill, A. "A Study of Thinking Mathematics" (1992). American Federation of Teachers, Washington, D.C.

Reprinted with permission from TEACHERS' LAB, November/December 1992,
Teachers' Laboratory, Inc., PO Box 6480, Brattleboro, VT 05302, 802-254-3457.

Bringing meaning to math with a student-run store

%

wholesale

retail

loan

inventory

$

¢

by Deborah Black

IMAGINE THE range of skills and concepts that can be addressed in the undertaking of a student-run math store. Skills include estimating simple sums and differences, multiplying with money, introducing decimals and many others. A 5% markup of wholesale items or a 10% discount on sale items offers a wonderful lesson on percentage!

Fractions enter the picture when you consider cutting a pie in eighths and selling individual pieces. Consider the complexity of a word problem that begins with the price of the ingredients in oatmeal cookies. How many cookies are in a batch? How much will you sell the individual cookies for if you want to account for a 5% profit? Keep in mind all the measuring involved in making the cookies and the importance of following the recipe directions. For some students, learning to make change for a dollar is a formidable task in itself.

Goals

The academic goals of our store were clear at the onset. The store would provide students with an opportunity to develop math skills with an improved attitude towards math. Students were continually engaged in problem solving at every level of store operation. Collegiality and cooperative group work began to unfold.

Fresh ideas

Each year the store has a fresh beginning as a new group of students bring to it their ideas and student-made merchandise. The only dimensions of the store that remain constant are the enthusiasm, excitement and ongoing learning. Students take charge of their own learning and begin to make connections between participating in different aspects of running the store and understanding certain math concepts and skills.

Preparing for the math store was the math curriculum for the initial weeks of school. Concepts and skills were reviewed or introduced as they related to some task in running the store. Peer teaching and apprenticing practices were developed in the first steps of the preparation process so that ongoing problem solving related to the store could take place in an atmosphere of support and collegiality. Later in the year students ran the store independent of other projects going on during math class.

Establishing respect and support

A lot of thought and consideration went into understanding what it meant to work together with a common purpose. Time was spent defining a mission for the store and a code of ethics for all participants. It was clearly stated in their code of ethics that all students had the right to participate in every aspect of running the business. It was made clear that it was the responsibility of the group to make sure everyone felt supported and treated with respect as they learned new skills necessary to run their business. Peer teaching and apprenticing were concepts that had been addressed earlier in the class. Once the foundation for working together was established, the process of preparing students to run their store began.

Defining jobs

A division of labor was established and a job description for each job was written by the students. Job titles included salesperson, cashier, buyer, inventory clerk, record keeper and even crafts people and bakers. Later on the students added to their list of jobs: loan officer, advertisement coordinator, and entertainment coordinator.

Each job description, as defined by the students, also stated what types of skills were necessary to be successful at each job. For example:

A salesperson needs to:

- have knowledge of merchandise in stock;
- be friendly and helpful to people, even when rushed;
- be able to add simple numbers in your head;
- know how to add the cost of merchandise to help the customer determine if he/she has enough money to pay for the item(s);
- know how to discount sale items.

These descriptions were always subject to change as students discovered new areas of expertise they felt were important to each job.

Practice run

Before the students actually put their store into operation, we had a practice week in math class. Each student brought in an item for under a dollar to sell at the "trial run." We went through every step of running a store that they would encounter: ordering merchandise, markup from wholesale to retail prices, opening and closing procedures for setting up the till, selling, making change, record keeping, keeping an inventory and re-ordering. Centers were set up so that every student was able to participate in all the aspects of running the store during practice week.

What to sell

Careful consideration should be given to this matter. We decided to sell school supplies, nutritional snacks and student-made crafts. We also decided to sell items that were under a dollar and put a dollar cap for spending. We wanted everyone to have an opportunity to participate in the shopping and did not want money to become a status issue.

Start-up costs & physical set-up

Students wrote a letter to the PTO asking for a loan of $65.00 to invest in stock with a payback schedule attached.

Once the loan came through students started preparing for their grand opening. Each student signed up for their first job. Those who were hesitant about their ability to perform certain tasks could be an apprentice to a peer teacher.

For a physical set-up, nothing elaborate is necessary. We used a roll-away cart that we actually rolled into different classrooms. Designing a math store structure would be a wonderful construction project!

The grand opening

The opening date was well advertised in the 5th/6th grades. With a bustle of energy and enthusiasm, the grand opening was a success. Over time, changes and additions to store routine became apparent and needed to be worked on collectively. Students took on more responsibility and began to organize baking groups and encourage student artists to sell their wares. They were motivated and challenged by the demands of their customers and their drive to make their store a success.

Regular follow-up and reflection

After every store closing, students would work together to prepare an update report for the group. At a group gathering students presented a financial report and an update on inventory items. A discussion followed that allowed students to reflect on the store operations and make suggestions for improvements and changes. Task forces were established to address issues brought up in the discussion and each student chose a job for the next opening of the store.

Students were continually engaged in problem solving at every level of store operation.

Evaluation and assessment

Noting the math skills and concepts covered by running different aspects of the store made it comfortable to assess student learning by watching them prepare and perform their job. Students also kept a math journal to reflect on the trials and tribulations of running the store. This served as a valuable tool for assessing and evaluating students learning. Attending to students' participation in problem solving related to the store was also valuable in assessing their development and understanding of math concepts and skills.

Extension and integration

Many opportunities for integration present themselves as students discover new ways to serve their customers and operate their store. For instance, students decided to sell only "healthy" snacks at our store which became a springboard for discussing nutrition. One year we had another class make jewelry and we sold their merchandise on commission in our store.

The store has been a wonderful opportunity for students to actively engage in hands-on math in a meaningful way. Each year as a new group of students comes to the class ready to begin their own business, the extent of involvement and complexity of the operation is determined by the collective needs of the students. The same process and structure for preparing to operate the store and the actual running of the store provide a skeleton from which I operate. However, the details and meat of the matter are dependent on the students and the extent to which they are developmentally ready for the math concepts and skills covered in the different aspects of running the store.

salesperson

cashier

buyer

loan officer

DEBORAH BLACK has a B.S. in Communications Disorders and a M.Ed. in Deaf Education and has taught for 10 years overseas and in the United States.

Reprinted with permission from TEACHERS' LAB, October 1992, Teachers' Laboratory, Inc., PO Box 6480, Brattleboro, VT 05302, 802-254-3457.

Employer expectations for school mathematics

by Henry O. Pollak

WHY DOES SOCIETY give us so much time to teach mathematics? There are, of course, reasons internal to the educational system: This is needed to pass the test, that is needed for the next course, etc. At this point, I am more interested in reasons outside of education itself. We need to teach mathematics

- for practical everyday life;
- for intelligent citizenship;
- for employment;
- as part of overall human culture.

Experience with problem formulation is exactly what we need in business and industry.

There is much overlap among these purposes. We will focus our attention on the mathematics needed for employment.

My views are based on 35 years of experience in the telephone industry, most of it in the Bell System prior to divestiture. As a mathematician, I had the job of sticking my nose into everybody's business. On the basis of this lifelong experience of myself and my many colleagues, here are some thoughts about mathematical expectations of employers:

1. Employees need to know how to set up a problem, that is, to take a situation in the real world of work and formulate a precise question the answer to which would help with the situation at hand.
2. They need to be able to figure out what an answer should look like before they start to solve. If the problem is numerical, for example, about how big should the answer be, and what sort of precision should they aim for?
3. Once the problem and a desired accuracy are defined, the employee should know, or know where to find, SOME method of solving it. What method is used is not important!
4. The employee should know that it is possible to look at a tremendous variety of practical situations in an analytical, structured, systematic, quantitative, i.e., "mathematical" way.
5. Employees need to be able to work together, to function as a team, to help each other.
6. They need to be able to communicate, to get across by the spoken and the written word what they are thinking about and what they have done.

These are simplified, abbreviated statements of need, but I think that business will recognize their importance.

Let me now switch hats — but still on the same head — and ask what these needs say to the processes and content of mathematics education. How is all this related to the *Standards*, and to the many current movements and changes? Here are some further thoughts:

1. A key difference between "NCTM's Agenda for Action," from 1980, and the *Standards* almost 10 years later, is the evolution from the prime emphasis on problem solving in 1980 to problem finding, before you try to solve, in 1990. Experience with problem formulation is exactly what we need in business and industry.
2. One effect on mathematics education of the rapidly changing technology is the even greater need than before for teaching estimation (at the developmentally appropriate level). This is also exactly what the employer needs. Before you start any problem, you ought to have an idea how big the answer ought to be. You also need to know, as was pointed out above, what precision is required. When you have done both of these, then you can consider alternative methods. If it is a numerical problem, do you want to do it in your head, or on a calculator, or by paper and pencil, or on a computer? It is terribly important that future employees have learned to do all four of these and have developed some (metamathematical) judgement about when to use which.
3. There is a great variety of mathematical thought. In the past, the bulk of students — i.e.

Expectations in the Workplace for School Mathematics

- ✔ The ability to set up problems with the appropriate operations.
- ✔ Knowledge of a variety of techniques to approach and work on problems.
- ✔ Understanding of the underlying mathematical features of a problem.
- ✔ The ability to work with others on problems.
- ✔ The ability to see the applicability of mathematical ideas to common and complex problems.
- ✔ Preparation for open problem situations since most real problems are not well formulated.
- ✔ Belief in the utility and value of mathematics.

(Pollak, 1987)

future employees — have at best seen three of these in school: geometric, arithmetic, and algebraic. In the world of work, we also meet situations which are inherently uncertain. The information is essentially **probabilistic**, and students must learn how to deal with it. There are other situations which are **data-driven**, and future employees need to know how to look at data and draw potential conclusions from them. Many real-world situations require planning, optimizing, choosing among alternatives. Thus the beginnings of the mathematics we call **operating research** are very useful. Many real-world situations are combinatorial or graph-theoretical or algorithmic in nature, and students should therefore have seen these kinds of mathematics.

4. The fact that it is possible to look at so much of the world in mathematical ways is the essence of mathematical modelling. Students need to have experienced this.

5. One of the most helpful developments in education in recent years is cooperative learning. When students learned only the behavior of competing with one another, one of the first, and most difficult, behaviors which employers have had to teach is cooperation. Group work will, we hope, help with this problem.

6. Language across the curriculum is another very hopeful and valuable trend. A piece of work is of no use to the company if it is neatly engraved in the employee's head or notebook, but unavailable to anyone else. Knowing how to speak and how to write is essential in making yourself useful to your employer. Writing in mathematics will help to prepare the future employee for this: so will oral reports of work.

The competitive position of the United States in the twenty-first century makes it essential that we do better in mathematics education than we have in the past. An awareness of the needs of employers will help to move the System in directions which will benefit everyone.

HENRY O. POLLAK is a Visiting Professor at Teachers College/Columbia University, and retired Assistant Vice President of Bell Communications Research, Inc.

Reprinted with permission from EDUCATIONAL LEADERSHIP, May 1992.

Evaluating Problem Solving in Mathematics

Effective assessment of problem solving in math requires more than a look at the answers students give. Teachers need to analyze their processes and get students to communicate their thinking.

WALTER SZETELA AND CYNTHIA NICOL

In its *Curriculum and Evaluation Standards for School Mathematics*, the National Council of Teachers of Mathematics expanded the goals it developed in 1980 for promoting problem solving as a curricular focus (NCTM 1989). The first three standards — Mathematics as Problem Solving, Mathematics as Reasoning, and Mathematics as Communication — show a shift from emphasis on rules and routine problem solving dominated by teacher talk and passive learning, to active student participation, in which reasoning and communicating are stressed.

These efforts are admirable, but they create new challenges, especially in assessment of these higher-level skills. Problem solving requires considerable thinking, but even when students are able, they are not inclined to communicate their thinking. Without such communication, how can we reliably assess students' efforts to solve problems? Before discussing how to improve communication and assessment, it is useful to clarify the notion of a problem and problem solving.

The Nature of Problems and Problem Solving

Problem solving is the process of confronting a novel situation, formulating connections between given facts, identifying the goal, and exploring possible strategies for reaching the goal. A problem, then, is a situation in which the individual initially does not know any algorithm or procedure that will guarantee solution of the problem, but the individual desires to solve it.

Success in problem solving depends upon metacognitive processes, as described by Garofalo and Lester (1985). The following list summarizes the typical sequence of actions for successful problem solving:

1. Obtain an appropriate representation of the problem situation.
2. Consider potentially appropriate strategies.
3. Select and implement a promising solution strategy.
4. Monitor the implementation with respect to problem conditions and goals.
5. Obtain and communicate the desired goals.
6. Evaluate the adequacy and reasonableness of the solution.
7. If the solution is judged faulty or inadequate, refine the problem representation and proceed with a new strategy or search for procedural or conceptual errors.

These metacognitive processes are difficult to assess, but assessment can be expedited by creating problem situations that facilitate students' communication of their thinking.

Difficulties in Assessment of Problem-Solving Performance

The difficulty of assessing complex processes necessary for solving problems is exacerbated by the failure of students to communicate clearly what they have done or what they are thinking. Students are prone to make calculations without explanations, and calculations alone often fail to reveal sufficiently the nature of the solver's work and thinking. It is not enough to

FIGURE 1

ANALYTIC SCALE FOR PROBLEM SOLVING

Understanding the problem
0 - No attempt
1 - Completely misinterprets the problem
2 - Misinterprets major part of the problem
3 - Misinterprets minor part of the problem
4 - Complete understanding of the problem

Solving the problem
0 - No attempt
1 - Totally inappropriate plan
2 - Partially correct procedure but with major fault
3 - Substantially correct procedure with minor omission or procedural error
4 - A plan that could lead to a correct solution with no arithmetic errors

Answering the problem
0 - No answer or wrong answer based upon an inappropriate plan
1 - Copying error; computational error; partial answer for problem with multiple answers; no answer statement; answer labeled incorrectly
2 - Correct solution

check for right and wrong answers or to use multiple-choice formats for assessment of problem solving. As Silver and Kilpatrick (1988) state:

> A reliance solely on the sleek efficiency of multiple-choice (and other short answer) formats will severely hinder efforts to help students develop a reflective and interrogatory stance toward their learning.

If we can devise methods for eliciting better communication of students' thinking, we can perform more effective assessment. Such assessment measures the quality of students' thinking. This information can help teachers design and implement instruction to promote greater success in problem solving and can help administrators evaluate programs and curriculums.

Assessment of Solved Problems

The most natural and common method for assessing performance in problem solving is to obtain general impressions about the quality of a solution while scanning students' work. These general impressions are strongly influenced by the "proximity of correctness" of the answer. As a result, good solutions with minor errors due to carelessness that alter the answer dramatically can receive undeservedly low scores. Scales are available that focus more attention on solution procedures, enabling teachers to obtain fairer and more reliable scores. For example, Charles, Lester, and O'Daffer (1987) devised a scale that assigns separate scores to each of three stages in problem solving: understanding the problem, solving the problem, and answering the question. Figure 1 shows a modification of their scale, with increased emphasis given to the understanding and solving stages (Wilson 1991).

The Charles, Lester, and O'Daffer scale and its modified forms are easy to use. An advantage of such a scale is that a teacher may focus on only one of the stages. For example, a teacher who is emphasizing strategy selection and implementation can assess each student's solving procedure irrespective of the answer.

The California Assessment Program (Pandey 1990) includes comprehensive descriptions of various levels of performance for specific problems. This is appropriate for large-scale assessment programs. However, the classroom teacher has little time to construct scales for individual problems. Teachers need assessment procedures and scales that they can modify or use intact for a wide range of problems.

Problem solving requires considerable thinking, but even when students are able, they are not inclined to communicate their thinking.

Categorizing Responses to Problems

Scales for assessment of problem solving can be designed without creating an evaluative threat to students. Such a system of scales was constructed for use in the 1990 British Columbia assessment of problem solving (Szetela 1991). Instead of scoring the solutions only, teachers analyze the responses to problems on the basis of four categories: answers, answer statements, strategy selection, and strategy implementation (see fig. 2). Teachers can use a single category to determine how well their students are addressing a particular aspect of solving problems. One focus might be on strategies used. Another might be directed toward answer statements. Incomplete statements that fail to include the units taught or important contextual information may serve as focal points for teachers in their subsequent instructional activities.

FIGURE 2

CATEGORIES OF RESPONSES IN SOLUTIONS TO PROBLEMS

Answer	Strategy Selected	Implementation
1. Blank	1. Number sentence	1. No work shown
2. Undetermined	2. Select operations and calculate	2. Identifies data only
3. Incorrect	3. Algebraic	3. Problem misinterpreted
4. Correct	4. Non-systematic list	4. Strategy not clear
	5. Systematic list	5. Strategy initiated (table, graph, list) but incomplete or poorly implemented
Statement	6. Guess and test	
	7. Draw diagram	6. Conditions or possibilities overlooked
1. No statement	8. Look for pattern	7. Multiple secondary errors
2. No context	9. Logical reasoning	8. A single secondary error
3. No units	10. Use simpler case	9. Appropriate and complete
4. None required	11. Work backwards	
5. Complete	12. Undetermined	

Students are prone to make calculations without explanations, and calculations alone often fail to reveal sufficiently the nature of the solver's work and thinking.

Promoting Greater Communication

To further enhance assessment, we need to devise problem situations and questions that encourage and motivate students to communicate and explain their thinking. Figure 3 shows one way to do this. An already solved problem with a significant error, combined with a set of relevant questions about the solution, facilitates communication. As with an unsolved problem, students must form a suitable representation of the problem. Instead of solving the problem themselves, however, they analyze the given solution. Finally, they reveal their thinking by answering the pertinent questions. Answers to these questions can provide more comprehensive insights about the student's thinking in problem situations than more typical problem formats, in which students may have various levels of success but fail to reveal their thinking.

Assessment of responses to the questions accompanying the already solved problem can be done in less time than it normally takes for teachers to plod through the usual wide range of solution procedures for a given problem. The main goal of the example in Figure 3 is to determine whether or not a student understands a problem situation well enough to recognize the incongruity of the given answer despite excellent implementation of a good plan, with the problem solver running awry only in the careless writing of the answer statement. Teachers can provide continuing experiences for students to critically analyze solutions and communicate their observations and responses to relevant questions. Such practice can help students engage in reasoning, evaluating, and communicating, and can enable teachers to assess these problem-solving processes more effectively.

Other forms of problems with questions to stimulate thinking and written communication include the following:

- Present a problem with all the facts and conditions, but have the students write an appropriate question, solve the completed problem, and write their perceptions about the adequacy of the solution.
- Present a problem and a partial solution. Have students complete the solution.
- Present a problem with facts unrelated to the question. Have students comment about the quality of the problem or revise the problem to remove the incongruity.
- Have students explain how they would solve a problem using only words, then solve the problem and construct a similar problem.
- After students solve a problem, have them write a new problem with a different context but preserving the original problem structure.
- Present a problem without numerals. Have students supply appropriate numerals, estimate answers, and solve the problem.

Teachers can assess the quality of each response by using a scale such as the following:

FIGURE 3

Example of Problem that Asks Students to Communicate Thinking

A bowl contains 10 pieces of fruit (apples and oranges).
Apples cost 5 cents each and oranges cost 10 cents each.
All together the fruit is worth 70 cents.
We want to find how many apples are in the bowl.
Kelly tried to solve the problem this way.

$10 \times 5 = 50$
$2 \times 10 = 20$
70

There were 30 apples in the bowl.

$8 \times 5 = 40$
$3 \times 10 = 30$
$4 \times 10 = 40$
$6 \times 5 = 30$

Try to follow Kelly's work and solution. Then answer the questions.

1. Is Kelly's way of solving the problem a good one?

Yes

Tell why you think it is or is not a good way.

because it will tell you the posible silution which is what you want but she didn't read carfully

2. Did Kelly get the right answer?

No

Explain why she did or did not.

because there are only ten items in the basket

1. No response or simplistic or irrelevant response.

2. A relevant response but of minor importance with respect to the question or problem.

3. A reflective and significant response but with an important omission or misconception.

4. A comprehensive, logical, and correct response to the question or problem.

These suggestions for assessment of problem solving have the potential to reveal much more than we currently know about students' thinking, their conceptions, their weaknesses, and their strengths. With better awareness about students' knowledge and thinking, teachers can plan more effective instruction, and the outcome is more likely to be better learning of higher-order skills essential to success in problem solving. □

References

Charles, R., F. K. Lester, Jr., and P. O'Daffer. (1987). *How to Evaluate Progress in Problem Solving.* Palo Alto, Calif.: Dale Seymour Publications.

Garofalo, J., and F. K. Lester, Jr. (1985). "Metacognition, Cognitive Monitoring, and Mathematical Performance." *Journal for Research in Mathematics Education* 16: 163-176.

Pandey, T. (1990). "Power Items and the Alignment of Curriculum Assessment." In *Assessing Higher Order Thinking in Mathematics,* edited by G. Kulm. Washington, D.C.: American Association for the Advancement of Science.

Silver, E. A., and J. Kilpatrick. (1988). "Testing Mathematical Problem Solving." In *The Teaching and Assessing of Mathematical Problem Solving,* edited by R. I. Charles and E. A. Silver. Reston, Va.: The National Council of Teachers of Mathematics.

Szetela, W. (1991). "Open-Ended Problems." In *The 1990 British Columbia Assessment of Mathematics: Final Report,* edited by D. Robitaille. Victoria, B.C.: Ministry of Education Assessment Branch.

Wilson, D. (1991). "Analysis of Students' Problem Solving." Unpublished paper.

Walter Szetela is an Associate Professor and **Cynthia Nicol** is a mathematics teacher and a Ph.D. candidate in mathematics education. They can be reached at The University of British Columbia, Faculty of Education, Department of Mathematics and Science Education, 2125 Main Mall, Vancouver, B.C., CANADA V6T 1Z5.